UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE QUIXADÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MESTRADO ACADÊMICO EM COMPUTAÇÃO

JOSÉ ALAN FIRMIANO ARAÚJO

BUSCA POR SIMILARIDADE DE BOLETINS DE OCORRÊNCIA VIA EMBEDDINGS: UM ESTUDO DE CASO

QUIXADÁ

2023

JOSÉ ALAN FIRMIANO ARAÚJO

BUSCA POR SIMILARIDADE DE BOLETINS DE OCORRÊNCIA VIA EMBEDDINGS: UM ESTUDO DE CASO

Dissertação apresentada ao Curso de Mestrado Acadêmico em Computação do Programa de Pós-Graduação em Computação do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em computação. Área de Concentração: Ciência da Computação

Orientadora: Profª. Dra. Ticiana Linhares Coelho da Silva

QUIXADÁ

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

A689b Araújo, José Alan Firmiano.
Busca por similaridade de boletins de ocorrência via embeddings : um estudo de caso / José Alan Firmiano Araújo. – 2023.
54 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Campus de Quixadá, Programa de Pós-Graduação em Computação, Quixadá, 2023.
Orientação: Profa. Dra. Ticiana Linhares Coelho da Silva.

1. busca por similaridade. 2. embedding de palavras. 3. embedding de sentenças. 4. relatórios policiais. I. Título.

CDD 005

JOSÉ ALAN FIRMIANO ARAÚJO

BUSCA POR SIMILARIDADE DE BOLETINS DE OCORRÊNCIA VIA EMBEDDINGS: UM ESTUDO DE CASO

Dissertação apresentada ao Curso de Mestrado Acadêmico em Computação do Programa de Pós-Graduação em Computação do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em computação. Área de Concentração: Ciência da Computação

Aprovada em: 25 de Julho de 2023

BANCA EXAMINADORA

Profª. Dra. Ticiana Linhares Coelho da Silva (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. José Antonio Fernandes de Macedo
Universidade Federal do Ceará (UFC)

Prof. Dr. Régis Pires Magalhães
Universidade Federal do Ceará (UFC)

Prof. Dr. Vinícius Monteiro
Universidade Federal do Ceará (UFC)

Profª. Dra. Atslands Rego da Rocha
Universidade Federal do Ceará (UFC)

Profª. Dra. Aline Marins Paes Carvalho
Universidade Federal Fluminense (UFF)

À minha esposa, Dálete Lima Firmiano Araújo, que me incentivou e me apoiou em todos os momentos dessa caminhada.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a Deus, que me guiou e me deu força durante todo o processo de pesquisa. Sua presença em minha vida foi essencial para a minha perseverança e o sucesso deste trabalho.

Também gostaria de agradecer a minha orientadora Dra. Ticiana Coelho da Silva, que me apoiou e me guiou durante todo o processo. Seus conselhos e sugestões foram inestimáveis para o sucesso deste trabalho.

Agradeço à minha família e amigos, que sempre me apoiaram e me deram força para superar os desafios. Sem a fé e o amor de Deus e de vocês, eu não teria chegado até aqui.

Agradeço também ao Dr. José Antonio Fernandes de Macedo, Dr. Régis Pires Magalhães e Dr. Vinícius Monteiro, que gentilmente me ajudaram com suas contribuições valiosas e críticas construtivas.

Agradeço aos meus colegas, por todas as discussões enriquecedoras e pela amizade que construímos ao longo desses anos.

Por fim, gostaria de expressar minha gratidão a todas as instituições que financiaram esta pesquisa e me proporcionaram a oportunidade de estudar e me desenvolver academicamente. Obrigado.

“A tecnologia é a ponte entre o presente e o futuro.”

(Stephen Hawking)

RESUMO

Vários crimes ocorrem diariamente, e o primeiro passo na investigação começa com um boletim de ocorrência. Em cidades com altos índices de criminalidade, é desafiador para a polícia lidar com a análise detalhada de todos os relatos criminais. No entanto, os boletins de ocorrência podem ser similares por apresentarem o mesmo *modus operandi*. Dado um boletim de ocorrência, o objetivo principal deste trabalho é determinar o mais similar ou duplicado. Um boletim de ocorrência similar pode ser outro relatório com palavras sobrepostas ou um que compartilhe um *modus operandi* similar. Uma solução possível é representar cada boletim de ocorrência como um vetor de caracteres e comparar os vetores usando uma função de similaridade. Diferentes métodos podem ser empregados para representar a narrativa, incluindo vetores de incorporação e abordagens baseadas em contagem, como o TF-IDF. Esta pesquisa explora o uso de representações de incorporação pré-treinadas nos níveis de palavra e sentença, como Universal Sentence Encoder, Word2Vec, BERT, Doc2Vec, entre outros. Determinamos a representação mais eficaz para capturar similaridades semânticas e léxicas entre relatórios policiais, comparando diferentes modelos de incorporação. Além disso, comparamos a eficácia de modelos de incorporação pré-treinados disponíveis com modelos treinados especificamente em um corpus de relatórios policiais. Outra contribuição deste trabalho é o desenvolvimento de modelos de incorporação treinados especificamente para o domínio de relatórios policiais.

Palavras-chave: busca por similaridade. embedding de palavras. embedding de sentenças. relatórios policiais.

ABSTRACT

Several crimes happen daily, and the first step in the investigation begins with a police report. In cities with high crime rates, it is challenging for the police to handle the detailed analysis of all criminal reports. However, incident reports may be similar as they present the same *modus operandi*. Given an incident report, the main objective of this work is to determine the most similar or duplicate. A similar police report may be another report with overlapping words or one that shares a similar *modus operandi*. One possible solution is to represent each police report as a vector of characters and compare the vectors using a similarity function. Different methods can be employed to represent the narrative, including embedding vectors and count-based approaches such as TF-IDF. This research explores the use of pre-trained embedding representations at both the word and sentence levels, such as Universal Sentence Encoder, Word2Vec, RoBERTa, Doc2Vec, among others. We determine the most effective representation for capturing semantic and lexical similarities between police reports by comparing different embedding models. Furthermore, we compare the effectiveness of available pre-trained embedding models with models specifically trained on a corpus of police reports. Another contribution of this work is the development of embedding models trained specifically for the domain of police reports.

Keywords: similarity search. word embedding. sentence embedding. police reports.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os boletins de ocorrência descrito pelas vítimas ou testemunhas são usados como um "pontapé inicial" para a investigação policial dos fatos denunciados. Um boletim de ocorrência também serve para resguardar a própria ação policial, demonstrando de onde teve início aquele conjunto de ações investigativas que o órgão policial está realizando. Em casos extremos, um boletim de ocorrência pode até chegar ao Supremo Tribunal Federal. Assim, um boletim de ocorrência deve ser detalhado e conter informações precisas sobre um incidente ou crime (por exemplo, factual, preciso, claro, conciso, completo e oportuno).

Em locais com alta ocorrência de crimes, diversos boletins de ocorrência são registrados todos os dias. No entanto, é desafiador para a polícia lidar com a análise de todos os boletins relatados. Na prática, existem vários problemas que os policiais enfrentam relacionados às tarefas de Processamento de Linguagem Natural (PLN): Reconhecimento de Entidade Nomeada (REN) para extrair as informações relevantes da narrativa como vítimas, o ladrão e as testemunhas. Outro problema relacionado pode ser a Classificação de Texto, ou seja, a partir da narrativa do boletim de ocorrência, categorizar a narrativa em um tipo de crime como roubo, fraude ou latrocínio, por exemplo. Outra tarefa da PLN enfrentada pelos policiais é identificar boletins policiais duplicados ou que relatem *modus operandi* semelhantes. Essas duplicações podem surgir devido à falta de conhecimento prévio sobre o incidente já reportado ou devido a problemas técnicos no sistema de registro. Em criminologia, é muito provável que o criminoso profissional siga sua maneira particular de operar, portanto, mantém o mesmo padrão, seguindo uma sequência de fatos e forma de agir. A recuperação de boletins de ocorrência duplicados ou que sigam *modus operandi* semelhante é uma tarefa de busca por similaridade e podem orientar a polícia sobre como deve agir.

Dado um boletim de ocorrência, o objetivo principal deste trabalho é identificar o boletim de ocorrência mais semelhante. Considere uma parte de dois boletins de ocorrência reais da polícia alegados por dois declarantes diferentes: o primeiro: "*Augusta Dias e seu irmão foram espancados e seus carros foram roubados no dia, local e hora mencionados acima. O declarante informa que o autor era uma pessoa conhecida como Fagner. As vítimas ficaram gravemente feridas no local do incidente. Sem mais nada a declarar.*" O segundo: "*Mario Dias e sua irmã tiveram seus carros roubados por uma pessoa chamada Fagner. Augusta e Mario ficaram gravemente feridos no local do incidente. Sem nada mais a declarar.*" [1] Ambos os

[1] Por razões de sigilo e preservação dos envolvidos, os nomes mencionados na narrativa são fictícios.

declarantes relatam o mesmo crime em diferentes boletins policiais. Portanto, o mesmo crime seria relatado duas vezes. A polícia precisa garantir que situações com o mesmo *modus operandi* sejam tratadas da mesma forma em todos os casos.

O problema de similaridade de texto determina se dois textos são semelhantes não apenas por similaridade lexical, mas também precisa considerar a similaridade semântica. Por exemplo, ambas as sentenças "*O carro de Augusta foi roubado ontem à noite*" e "*Na noite passada, Augusta teve seu carro roubado*" têm o mesmo significado. Para comparar a similaridade de duas sentenças, uma ideia é representar cada boletim de ocorrência como um vetor numérico e comparar esses vetores por meio de funções, como distância euclidiana ou similaridade cosseno. Embeddings pré-treinados para representar palavras ou textos são uma das representações mais populares do vocabulário de documentos.

Existem vários embeddings de palavras pré-treinados (Mikolov *et al.*, 2013; Pennington *et al.*, 2014). Embora a representação de palavras seja capaz de captar o contexto de uma palavra em um documento, semelhança sintática, relação com outras palavras, não é eficaz para captar a mudança de significado caso ocorram pequenas mudanças em uma frase. Por exemplo, as frases "*o carro da Augusta foi roubado*"e "*o carro da Augusta não foi roubado*"têm significados opostos apesar de possuírem apenas uma palavra diferente. Mas, apesar da natureza semanticamente oposta, a similaridade do cosseno pode ser presumida como muito alta entre os vetores obtidos a partir dessas sentenças usando embeddings de palavras. A partir do modelo Word2Vec (Mikolov *et al.*, 2013), a similaridade cosseno entre as sentenças foi de 0.9604, mesmo sendo opostas, usando o pré-treinado [2]. Os métodos de embedding devem ser capazes de contornar esse problema e fornecer representações que sejam melhores de diferenciar dois textos.

Uma alternativa é codificar frases ou sentenças em vetores de embedding. Esta abordagem é chamada de embedding de sentenças e existem várias opções pré-treinadas, como o Universal Sentence Encoder (Reimers; Gurevych, 2019), SBert (Cer *et al.*, 2018) e Doc2Vec (Le; Mikolov, 2014). Basicamente, os modelos recebem como entrada uma lista de frases e produzem como saída um vetor de dimensões fixa para a representação da sentença. As técnicas de embedding de sentença representam frases inteiras e suas informações semânticas como vetores. Isso ajuda a máquina a compreender o contexto, a intenção e outras nuances de todo o texto. O principal fator que afeta os vetores de embedding de sentenças são os dados de

[2] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

treinamento fornecidos. As sentenças no conjunto de treinamento precisam ser semanticamente relacionadas para alcançar melhores resultados (Kiros *et al.*, 2015).

Este trabalho se concentra em estudar buscas por similaridade em documentos de boletim de ocorrência, tendo em vista as características especificas deste tipo de documento. Boletins de ocorrência se distinguem de outros domínios devido à utilização de uma linguagem técnica especializada própria da área jurídica, o que torna a tarefa de buscar por similaridades mais desafiadora. Além disso, esses documentos contêm informações detalhadas sobre o contexto dos incidentes, como localização, horário, pessoas envolvidas e eventos prévios, o que requer abordagens que levem em conta o contexto para uma análise precisa. Outro aspecto que os diferencia é a presença de informações sensíveis e confidenciais, como detalhes pessoais e materiais, o que exige precauções especiais para garantir a privacidade e cumprir as diretrizes legais e éticas. Portanto, é necessário adotar abordagens adaptadas e personalizadas para realizar uma análise eficaz da similaridade entre esses boletins de ocorrência.

O objetivo principal é encontrar boletins policiais semelhantes - com palavras sobrepostas ou boletins que compartilham o mesmo *modus operandi*. Para orientar esta pesquisa, formulamos as seguintes questões de pesquisa:

- **QP1.** Qual representação pré-treinada em nível de boletim captura melhor as características sintáticas do vocabulário usado em boletins de ocorrência e pode distinguir entre boletins policiais duplicados? Para determinar a representação ideal da frase de entrada, este estudo leva em consideração os embeddings de palavras e frases, bem como técnicas de representação de texto baseadas em contagem, como TF-IDF;
- **QP2.** Qual é a melhor representação de embedding pre-treinada quando os boletins de ocorrência apresentarem o mesmo (*modus operandi*), não necessariamente sobreposição de palavras? Em outras palavras, qual é a melhor representação de boletim que captura a similaridade semântica? Por exemplo, dois boletins policiais diferentes de um roubo ocorreram com duas mulheres diferentes, mas em um caso, o ladrão roubou o celular e estava em uma moto com uma arma. Já na outra, o ladrão também estava em uma moto, armado com uma faca, mas roubou a bolsa da vítima;
- **QP3.** Quão eficazes são os modelos de embeddings para representação, quando utilizam treinamento continuado com documentos de boletins policiais? Nosso objetivo é analisar se esses modelos são melhores em representar boletins para buscas por similaridade. Além disso, comparamos aos modelos pré-treinados usados em **QP1**. Nosso objetivo principal

é avaliar o desempenho dessas várias versões (pré-treinada e treinada usando dados do domínio) quando se trata do problema de busca por boletins policiais semelhantes.

– **QP4.** A combinação de embeddings poderia alavancar a qualidade para busca por boletins policiais semelhantes? Com base em (Ghannay *et al.*, 2016), a combinação de diferentes embeddings tira proveito de sua complementaridade e produz uma melhoria em diferentes tarefas. Tomamos as duas melhores incorporações da análise das questões de pesquisa anteriores e comparamos a eficácia da representação de dois modelos, usando o vetor máximo, mínimo e médio. Esta questão de pesquisa investiga se é possível obter uma melhor representação dos boletins mesclando representações de dois modelos.

1.1 Contribuições

Este trabalho apresenta contribuições significativas, destacando-se as seguintes:

– Identificação do modelo pré-treinado mais adequado para a detecção de boletins de ocorrência duplicados. Foram realizados experimentos e análises comparativas entre diferentes modelos pré-treinados: Flair, Universal Sentence Encoder (USE), Doc2Vec, Word2Vec, BERT, Sbert e GPT, visando identificar aquele que apresenta o melhor desempenho na tarefa de identificação de boletins de ocorrência similares.

– Desenvolvimento de modelos de representação de sentenças que foram treinados levando em consideração o contexto dos boletins policiais, como: Word2Vec, Doc2Vec e RoBERTa. Esses modelos foram adaptados e ajustados para lidar com a linguagem e as características específicas dos boletins de ocorrência, buscando capturar com precisão as nuances e informações relevantes presentes nesses documentos.

– Estudo sobre a combinação de diferentes modelos de representação de sentenças para a representação mais adequada de boletins de ocorrência. Foram exploradas abordagens que envolvem a utilização conjunta de múltiplos modelos de embeddings: Word2Vec + Universal Sentence Encoder, RoBERTa + Universal Sentence Encoder. Permitindo obter uma representação mais abrangente e rica das sentenças presentes nos boletins de ocorrência, contribuindo para uma identificação mais precisa e confiável de sentenças duplicadas.

Essas contribuições visam aprimorar os métodos e técnicas utilizados na área de Processamento de Linguagem Natural para a análise e comparação de boletins de ocorrência, proporcionando resultados mais precisos e relevantes para a identificação de boletins semelhantes

e auxiliando na investigação de crimes.

1.2 Publicações e Submissões

Durante o desenvolvimento deste trabalho, ocorreu a publicação do artigo *Identifying Duplicate Police Reports* publicado na *International Conference on Machine Learning and Applications (ICMLA)* em 2021, Qualis A2. O referido artigo apresenta uma abordagem para identificar boletins policiais duplicados ou semelhantes, através da experimentação de diferentes representações pré-treinadas em nível de sentença para capturar as propriedades semânticas do vocabulário dos boletins. Além disso, investigou-se o efeito da sumarização dos boletins policiais na capacidade de identificar boletins duplicados. Alguns modelos e métodos de similaridade presentes na atual pesquisa foram validados nesta publicação.

No momento da escrita dessa dissertação, o artigo intitulado *Police Report Similarity Search: a case study* foi submetido e aceito para publicação no *Brazilian Conference on Intelligent Systems (BRACIS)* 2023, Qualis A4, onde os experimentos e resultados desta pesquisa são apresentados.

1.3 Estrutura da Dissertação

Este trabalho apresenta a seguinte organização: O Capítulo 2 aborda os conceitos e fundamentos teóricos essenciais para o entendimento do tema, como a natureza dos boletins policiais, os desafios relacionados à identificação de boletins duplicados e similaridades, e as técnicas de processamento de linguagem natural utilizadas no estudo.

Em seguida, o Capítulo 3 discute as pesquisas anteriores e estudos relacionados ao tema em questão. São apresentadas as abordagens existentes, os métodos utilizados e as principais contribuições dos trabalhos relacionados.

No Capítulo 4, descreve-se em detalhes a metodologia adotada neste trabalho. São apresentados os passos seguidos, incluindo a coleta e preparação dos dados, a seleção das técnicas de representação e comparação dos boletins, bem como as métricas utilizadas para avaliar o desempenho dos modelos avaliados.

O Capítulo 5 apresenta os resultados obtidos a partir da aplicação da metodologia proposta. São apresentadas as análises dos experimentos realizados e as descobertas obtidas com base nos resultados.

Por fim, o Capítulo 6 apresenta as conclusões do trabalho, destacando as principais contribuições, limitações e possíveis melhorias. Também são discutidos os caminhos para trabalhos futuros, indicando as áreas que podem ser exploradas para aprimorar o sistema de identificação de boletins policiais duplicados.

2 FUNDAMENTAÇÃO TEÓRICA

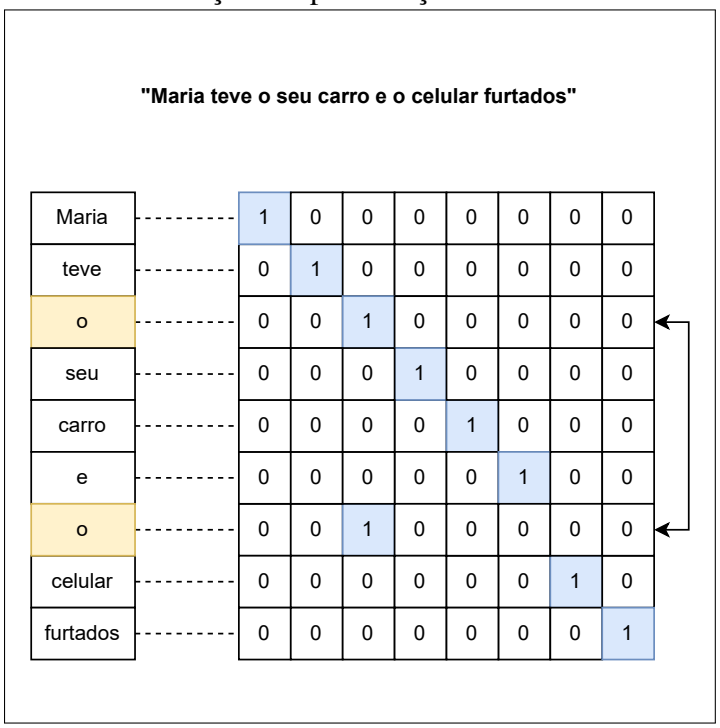
Nesta seção, apresentamos a fundamentação teórica, com os principais conceitos relacionados ao nosso problema. A Seção 2.1 apresenta os conceitos iniciais sobre a representação de sentenças como vetores. Nessa parte, são discutidos os modelos utilizados e a forma como eles representam as frases em vetores numéricos, tornando-as assim passíveis de serem processadas por algoritmos de aprendizado de máquina. A Seção 2.2 trata da representação vetorial de palavras. É explicado o conceito de embedding de palavras, que consiste na representação de palavras como vetores numéricos em um espaço de alta dimensionalidade. São apresentados alguns modelos de embedding de palavras, como Word2Vec e GloVe. A Seção 2.3 discute a representação de sentenças e documentos, incluindo a variação do contexto. Nessa parte, são apresentados modelos de embedding de documentos, como o Doc2Vec e o Universal Sentence Encoder (USE), que capturam o contexto das sentenças e documentos em que as palavras estão inseridas.

2.1 Representação de Textos

O método de vetorização de textos envolve transformar uma palavra ou uma sentença em vetor numérico. A representação mais frequente e fundamental é a codificação *one-hot*. Cada palavra recebe um índice inteiro distinto, e esse índice inteiro i é então convertido em um vetor binário de tamanho N (tamanho do vocabulário). Todos os elementos do vetor são zeros em sua composição, exceto para a i-ésima entrada, que é 1 (Chollet, 2021). A Figura 1 mostra um exemplo de como a sentença *"Maria teve o seu carro e o celular furtados"* seria representada nessa codificação. Observe que para cada palavra no vocabulário, existe uma representação única, ou seja, o artigo *"o"* mesmo aparecendo duas vezes na sentença contém a mesma representação.

Outra representação comum é usar *bag-of-words* ou *bag-of-n-words* (Harris, 1954). De acordo com esse paradigma, uma frase ou texto é tipicamente representado como uma coleção de palavras, ignorando a gramática e a ordem das palavras, mas mantendo a multiplicidade. Na Figura 2 podemos ver a mesma frase: *"Maria teve o seu carro e o celular furtados"*, porém, nessa representação, cada elemento do vetor representa a contagem de ocorrências de uma palavra específica no vocabulário. No entanto, essas técnicas apresentam algumas desvantagens, como vetores com alta dimensionalidade. Perceba que nos exemplos não removemos algumas palavras irrelevantes, tal como o artigo ("o").

Figura 1 – Exemplo de uma sentença na representação one-hot



Fonte: elaborado pelo autor (2023).

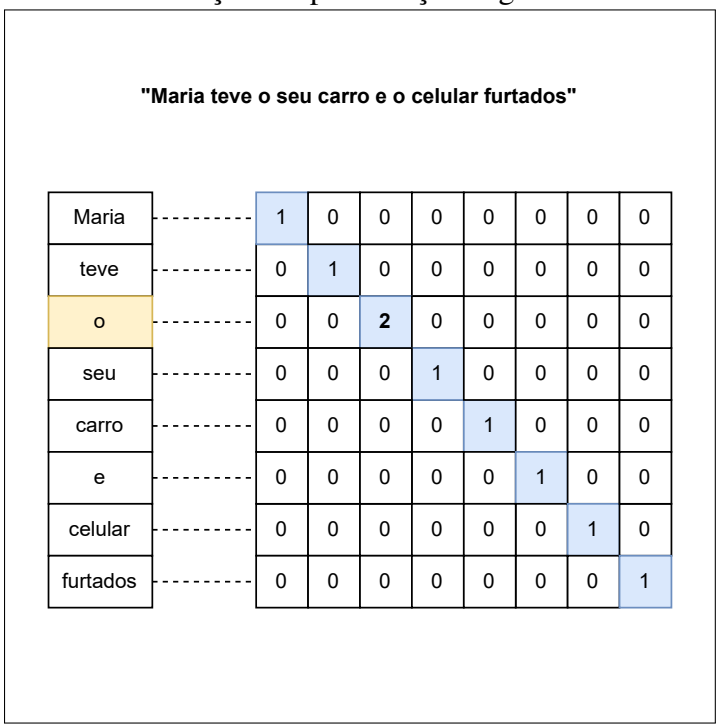
Existem algumas alternativas para representar cada palavra em um documento vetorialmente: (i) um booleano indicando a presença ou não; (ii) a frequência da palavra no texto; (iii) TF-IDF (*Term Frequency, Inverse Document Frequency*), uma medida estatística que avalia a importância de uma palavra para um documento dentro de uma coleção de documentos. O TF-IDF será o produto obtido das duas métricas:

- TF (Frequência do Termo): mede a frequência com que uma palavra aparece em um documento. Quanto mais frequente uma palavra ocorre em um documento, maior é o seu valor de *Term Frequency*.

- IDF (Frequência Inversa do Documento): intuitivamente, mede a raridade de uma palavra em um conjunto de documentos. Palavras que aparecem em muitos documentos terão um valor de *Inverse Document Frequency* menor, enquanto palavras raras terão um valor maior.

No entanto, essas técnicas apresentam algumas desvantagens, como vetores de alta dimensionalidade e dificuldade em identificar casos de polissemia, onde uma palavra pode ter vários significados. Neste trabalho, experimentamos, além das representações via embeddings de sentença, a representação de cada palavra em uma frase como seu valor para a medida TF-IDF.

Figura 2 – Exemplo de uma sentença na representação bag-of-words



Fonte: elaborado pelo autor (2023).

2.2 Embeddings de Palavras

Os modelos de espaço vetorial transformam o texto de diferentes comprimentos (como uma palavra, frase, parágrafo ou documento) em vetores numéricos de tamanho fixo, que são então utilizados em etapas subsequentes (como identificação de similaridade ou modelos de aprendizado de máquina). Embeddings de palavras pré-treinados têm sido amplamente usados (Mikolov *et al.*, 2013; Pennington *et al.*, 2014; Akbik *et al.*, 2019), devido à sua capacidade de capturar o contexto de uma palavra em um documento, semelhança semântica e sintática com outras palavras.

O artigo (Mikolov *et al.*, 2013) propõe um framework para aprender os vetores de palavras, treinando um modelo de linguagem que prevê uma palavra dada as outras palavras em um contexto. Uma implementação particular de tal estrutura é o Word2Vec. A principal desvantagem é que (Mikolov *et al.*, 2013) utiliza mal as estatísticas do corpus, uma vez que o modelo é treinado em uma janela de contexto local separada ao invés de contagens globais de co-ocorrência. (Pennington *et al.*, 2014) contorna este problema e propõe Glove, um modelo que produz um espaço vetorial de palavras. (Pennington *et al.*, 2014) treina o modelo em contagens globais de co-ocorrência palavra-palavra e faz uso eficiente de estatísticas. Outra proposta de representação de palavras é o Flair (Akbik *et al.*, 2019), que abstrai os desafios de engenharia

específicos que diferentes tipos de embeddings de palavras agregam, criando uma interface unificada para todos os embeddings de palavras e frases, bem como combinações arbitrárias de embeddings.

Os modelos BERT treinados para a língua portuguesa (Brasil), em específico, foi introduzido pelo BERTimbau (Souza *et al.*, 2020). Foram usadas três tarefas de processamento de linguagem natural para avaliar os modelos - similaridade textual da frase, reconhecimento de ligação textual e reconhecimento de entidade nomeada. Comparado ao BERT multilíngue e às técnicas anteriores de linguagem única, o BERTimbau é o que há de mais moderno nessas tarefas, demonstrando o valor de grandes modelos de linguagem pré-treinados para o português.

A partir de embeddings de palavras, podemos obter vetores de documentos. Uma alternativa é fazer a média de todos os vetores de palavras. No entanto, esse procedimento dá o mesmo peso a palavras importantes e não importantes. Outra limitação de representar texto usando embeddings de palavras é que cada palavra seria incorporada com o mesmo vetor, independentemente do contexto. Considere a frase "*Paris Hilton viajou dos Estados Unidos para Paris*". Por exemplo, a palavra *Paris* dependendo do contexto indica que pode estar se referindo ao nome (por exemplo, Paris Hilton) ou uma localização. Uma extensão dos embeddings de palavras são os embeddings de documentos, utilizados para obter os vetores diretamente do documento. Discutimos tais propostas na próxima seção.

2.3 Embedding de Documentos

A representação de documentos é realizada em um espaço vetorial n-dimensional de forma que palavras semanticamente semelhantes ou semanticamente relacionadas se reúnam no método de treinamento. Além disso, com base no contexto em que esteja sendo empregado uma palavra terá diferentes representações.

Existem muitas propostas para embeddings de frases como InferSent (Conneau *et al.*, 2017), LaBSE (Feng *et al.*, 2020), Universal Sentence Encoder (Reimers; Gurevych, 2019), Doc2Vec (Le; Mikolov, 2014), SBert (Cer *et al.*, 2018), entre outros. Um dos modelos mais famosos para embedding de sentenças foi o GPT Embedding da *OpenAI* (Neelakantan Tao Xu, 2022). O GPT Embedding é um modelo de linguagem treinado em uma tarefa de previsão de palavras seguintes (ou seja, tarefa de autoregressão) (Neelakantan Tao Xu, 2022). A etapa de treinamento do modelo foi realizada de forma unidirecional, ou seja, ele prevê palavras apenas com base no contexto anterior, e de forma não supervisionada (Neelakantan Tao Xu, 2022).

O Universal Sentence Encoder (USE) é um modelo que oferece duas opções diferentes para codificação de frases. Utilizando arquitetura Transformer (Vaswani *et al.*, 2017), o USE obteve o melhor desempenho em comparação com outros modelos de embeddings. O mecanismo de atenção calcula representações sensíveis ao contexto de palavras em uma frase que leva em consideração a ordem e a identidade de todas as outras palavras. As representações de palavras sensíveis ao contexto são convertidas em um vetor de codificação de sentenças de comprimento fixo calculando a soma dos elementos das representações em cada posição da palavra (Reimers; Gurevych, 2019). O outro codificador proposto é baseado em uma rede de média profunda (DAN) (Iyyer *et al.*, 2015) em que embeddings de palavras de entrada e bi-gramas são calculados juntos e depois passados por uma rede neural profunda *feed-forward* para produzir embeddings de frases.

Semelhante ao Word2Vec, o modelo Doc2Vec treina os vetores de parágrafo (ou embeddings de sentença) na tarefa de previsão da próxima palavra, levando em consideração o contexto que está inserido na frase. O vetor de parágrafo ou da sentença e os vetores de palavras são concatenados para prever a próxima palavra em um contexto. SBert (Cer *et al.*, 2018) pega a rede BERT pré-treinada e adiciona uma operação de agrupamento à saída para derivar uma representação de frase fixa. A fim de ajustar o BERT, o SBERT apresenta as estruturas de rede siamesa e tripla para derivar embeddings de frases semanticamente significativas que podem ser comparadas usando similaridade cosseno. Em (Cer *et al.*, 2018), os autores experimentam algumas estratégias de agrupamento: usando a saída do CLS-token, para calcular a média de todos os vetores de saída (estratégia padrão) e calcular um máximo ao longo do tempo do vetor de saída.

O modelo LaBSE treinado é proposto por (Feng *et al.*, 2020) para representação de sentenças em vários idiomas. Apenas pares de expressões bilíngues que são traduções umas das outras são usados exclusivamente no treinamento e otimização do LaBSE para gerar representações comparáveis. As frases de origem e destino são codificadas independentemente usando um BERT-based encoder compartilhado no framework do LaBSE, em seguida, passadas para uma função de combinação. Isso é conhecido como codificador duplo. A representação da frase para cada entrada é obtida a partir das representações da última camada [CLS]. Para pontuar a similaridade entre as frases de origem e destino é utilizado a distância cosseno na representação da frase criada pelos codificadores BERT.

Neste trabalho, experimentamos uma variedade de modelos de representação, in-

cluindo Word2Vec, Flair, BERT, Doc2Vec, Universal Sentence Encoder e SBert.

2.4 Redes Neurais Siamesas

As redes neurais siamesas, apresentadas por (Bromley Isabelle Guyon; Shah, 1994), são uma arquitetura composta por duas redes neurais que compartilham os mesmos pesos, conectadas por uma ou mais camadas. Essa arquitetura é comumente utilizada para codificar dados de entrada de forma não linear, com o objetivo de mapeá-los para um espaço onde padrões semelhantes estejam próximos e padrões não relacionados estejam distantes (Harandi *et al.*, 2017).

As redes neurais siamesas têm sido amplamente utilizadas para resolver diversos problemas que envolvem a avaliação da similaridade entre dois objetos, especialmente no reconhecimento de similaridade entre imagens. Seu uso original foi na verificação de assinaturas para determinar se pertencem à mesma pessoa (Bromley Isabelle Guyon; Shah, 1994), e desde então tem sido aplicada em reconhecimento e/ou verificação de faces, reconhecimento de voz, entre outras aplicações citadas por (Harandi *et al.*, 2017). No campo de aprendizado, as redes neurais siamesas têm sido empregadas com sucesso como medida de similaridade (Neculoiu *et al.*, 2016).

No treinamento e teste, as redes neurais siamesas recebem pares de entradas, buscando estabelecer similaridade entre pares da mesma classe (classe genuína) e distanciar pares de classes diferentes (classe impostora). Durante o treinamento, a rede cria um espaço multidimensional, com dimensão igual ao número de neurônios de saída da rede base, compartilhada entre os pares de entrada. Em seguida, a representação espacial de cada entrada é submetida a uma função de similaridade, geralmente uma medida de distância Euclidiana.

Após calcular a medida de similaridade entre os dados de entrada, é necessário determinar se eles pertencem à mesma classe ou não, com base em um limite de distância estabelecido. Para isso, as redes neurais siamesas utilizam a Perda Contrastiva como sua função de custo durante o treinamento. Essa função de perda, introduzida por (Chopra *et al.*, 2005), permitiu um treinamento mais eficaz dessa arquitetura.

A Perda Contrastiva é calculada somando as perdas individuais para os pares genuínos e impostores. Durante cada época de treinamento, os pares genuínos são atraídos para um espaço próximo, enquanto os pares impostores são mantidos a uma distância maior que uma margem definida.

Essa abordagem de treinamento permite que a rede neural siamesa aprenda a distinguir entre pares de dados pertencentes à mesma classe (genuínos) e pares de classes diferentes (impostores), ajustando seus pesos de forma a minimizar a perda contrastiva e melhorar a capacidade de classificação e reconhecimento.

2.5 Mean Reciprocal Rank (MRR)

A avaliação de modelos de classificação e ranqueamento desempenha um papel crucial no desenvolvimento e aprimoramento de ferramentas de processamento de linguagem natural. Entre as diversas métricas disponíveis para essa finalidade, o *Mean Reciprocal Rank* (MRR), ou Classificação Recíproca Média, destaca-se como uma medida estatística amplamente reconhecida e utilizada.

O MRR é projetado para avaliar a capacidade de um sistema em apresentar os itens corretos em uma lista de respostas, ordenados por sua probabilidade de correção. Sua formulação estatística considera não apenas a presença do item correto na lista, mas também sua posição relativa, calculando a média das inversas das posições dos itens corretamente classificados para uma série de consultas, dado pela fórmula:

$$\text{MRR} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{1}{\text{rank}_i},$$

onde D é o conjunto de dados de boletins policiais e rank_i refere-se à posição de classificação do primeiro documento relevante (mais semelhante) para o i-ésimo boletim policial.

2.6 Medidas de Similaridade

As medidas que são amplamente empregadas sempre que duas ou mais representações vetoriais precisam ser comparadas para determinar o grau de similaridade são: Similaridade cosseno e o coeficiente de Jaccard. A similaridade do cosseno é formalizada da seguinte forma:

$$\cos(w_1, w_2) = \frac{w_1 \cdot w_2}{\|w_1\|\|w_2\|} = \frac{\sum_{i=1}^{n} w_{1i} w_{2i}}{\sqrt{\sum_{i=1}^{n} (w_{1i})^2}\sqrt{\sum_{i=1}^{n} (w_{2i})^2}},$$

onde w_1 e w_2 são os vetores de sentença a serem comparados. A similaridade cosseno determina a proximidade das duas sentenças calculando o produto escalar entre elas dividido por suas normas.

O coeficiente de Jaccard mede a semelhança entre as frases de acordo com as palavras que elas têm em comum. O coeficiente de Jaccard é definido como o tamanho da interseção (o número de palavras que ambas as frases têm em comum) dividido pelo tamanho da união (número de palavras diferentes que unem as duas frases):

$$\text{Jaccard}(w_A, w_B) = \frac{|w_A \cap w_B|}{|w_A \cup w_B|},$$

onde w_A e w_B representam o conjunto de palavras usadas nas sentenças A e B, respectivamente.

3 TRABALHOS RELACIONADOS

Este capítulo de Trabalhos Relacionados apresenta uma revisão bibliográfica dos estudos e pesquisas que foram realizados anteriormente sobre o mesmo tema abordado nesta dissertação. Neste Capítulo, destacamos os trabalhos mais relevantes relacionados ao campo de pesquisa da identificação de sentenças similares.

Ao comparar as metodologias e resultados desses estudos com a nossa pesquisa, é possível identificar as principais lacunas e oportunidades de contribuição para a área de identificação de sentenças ou documentos similares.

3.1 Estudo sobre similaridade entre textos

Vários trabalhos comparam diferentes embeddings no que diz respeito à sua eficácia para a semelhança de texto. (Rodrigues; Marcacini, 2022) abordaram a tarefa de similaridade textual em língua portuguesa, utilizando modelos pré-treinados com a arquitetura de redes siamesas do SentenceBERT. A similaridade entre pares de frases foi avaliada por meio da distância cosseno. O estudo empregou um conjunto de dados que abrangeu níveis de similaridade variando de 1 a 5, sendo 1 indicativo de nenhuma semelhança e 5 representando alta semelhança semântica entre as representações. Além disso, (Rodrigues; Marcacini, 2022) investigou a possibilidade de obter representações aprimoradas ao combinar modelos pré-treinados na identificação de sentenças semanticamente similares.

No artigo (Shahmirzadi *et al.*, 2019), os autores propõem um estudo comparativo da similaridade semântica entre diferentes métodos no espaço de patentes. Os espaços vetoriais usados em (Shahmirzadi *et al.*, 2019) são uma linha de base TF-IDF simples, modelo de tópico LSI (*Latent Semantic Indexing*) e modelo neural Doc2Vec. Já o artigo (Naili *et al.*, 2017), também utiliza modelagem de tópicos e diferentes representações de vetores de aprendizagem, mas em outro tipo de atividade, tais como *Latent Semantic Analysis* (LSA), Word2Vec e GloVe para determinar qual é o método mais eficiente no campo da segmentação de tópicos.

(Imtiaz *et al.*, 2020) aborda o problema de identificação de questões duplicadas. A abordagem vetoriza as perguntas com base na representação de vetor de notícias do Google, representação de rastreamento FastText e de sub palavras de rastreamento FastText, tanto individualmente quanto como uma combinação. Esses vetores de palavras das perguntas são usados para descobrir a semelhança semântica das palavras. O modelo proposto é treinado em todos os

recursos extraídos dessas representações de três palavras separadamente e passado para a Rede Neural MaLSTM que utiliza a *Manhattan distance* para determinar a similaridade semântica entre as questões. (Firmiano; Silva, 2021) trata de uma questão correlata. Comparando o desempenho da representação por vetores de palavras e vetores de sentenças a fim de detectar boletins policiais duplicados. (Firmiano; Silva, 2021) também investiga a utilização de embeddings de sentença para identificar boletins policiais duplicados, explorando se uma sentença sumarizada é capaz de manter a semântica da representação da sentença original.

Outros trabalhos empregam embedding de sentenças para mapear a entrada em vetores de baixa dimensão e realizar pesquisas. O trabalho de (Hjaltason; Samet, 2003) investiga como incorporar objetos de dados complexos (como sequências de DNA, imagens, etc) em um espaço vetorial para que suas distâncias sejam tão próximas quanto possível de suas distâncias reais. Como resultado, em um cenário de aplicação de pesquisa por similaridade, as consultas podem ser executadas nos objetos representados. (Do; Pham, 2021) faz uma pesquisa de similaridade em *Knowledge Graphs* (KG) e ajuda os usuários a identificar as entidades mais significativas para sua consulta. A representação de texto de KG é um subcampo do aprendizado de representação de rede (NRL) e ajuda a preservar e representar a estrutura de KG (entidades e relações) em vetores de baixa dimensão. (Do; Pham, 2021) resolve tarefas de pesquisa semelhantes em um gráfico de conhecimento por meio de representação. Em seguida, (Do; Pham, 2021) mede as entidades relevantes calculando a distância entre seus vetores de imersão.

3.2 Similaridade em palavras

Um problema relacionado a identificação de sentenças semelhantes pode ser encontrar palavras semelhantes. Existem alguns conjuntos de dados em inglês para tarefas de similaridade de palavras como MC (Miller; Charles, 1991), SimLex-999 (Hill *et al.*, 2015), RG (Rubenstein; Goodenough, 1965), WordSim353 (Finkelstein *et al.*, 2001), entre muitos outros. Alguns deles também foram liberados para associação de palavras como WordSim353 (Finkelstein *et al.*, 2001) e SimLex-999 (Hill *et al.*, 2015).

Associação de palavras e similaridade são duas tarefas diferentes, por exemplo [casa, apartamento] e [casa, cama]. Diz-se que a casa é semanticamente semelhante ao apartamento e associada (mas não semelhante) à cama. Casa e apartamento podem ser entendidos como semelhantes devido à sua função padrão (geralmente onde as pessoas moram) ou porque compartilham características físicas (geralmente, ambos possuem um quarto e uma cozinha). Por outro

lado, casa e cama estão associadas porque se espera que apareçam juntas nos documentos.

A tarefa de similaridade de palavras é comumente usada para avaliar a qualidade da representação de texto conforme avaliado nos artigos de Word2Vec (Mikolov *et al.*, 2013) e Glove (Pennington *et al.*, 2014). Essa tarefa é realizada como uma avaliação intrínseca desses trabalhos. Conforme apresentado em (Navigli; Martelli, 2019), a similaridade de palavras pode ser alcançada explorando os recursos de conhecimento léxico-semântico ou explorando a similaridade distributiva, ou seja, a distribuição estatística de palavras dentro do texto não estruturado.

A avaliação intrínseca é independente de um tarefa específica de Processamento de Linguagem Natural, avaliando diretamente as relações sintáticas ou semânticas entre as palavras. Os conjuntos de dados de referência criados e citados acima por humanos são gerados com um estudo específico de comparação de palavras por similaridade e parentesco, que são cruciais no processo de comparação da avaliação intrínseca. O método extrínseco de vetores de palavras é a avaliação integrada em uma tarefa de Processamento de Linguagem Natural como predição da próxima palavra ou análise de sentimento, escolhida como método de avaliação. Normalmente, as avaliações de embeddings de palavras na avaliação extrínseca reportam precisão e F1-score, entre outras métricas. Para mais detalhes, o leitor pode consultar (Zhai *et al.*, 2016; Qiu *et al.*, 2018).

Embora existam vários conjuntos de dados em inglês para similaridade de palavras, conforme citado anteriormente, conjuntos de dados em idiomas de poucos recursos, como português e japonês, são escassos. (Inohara; Utsumi, 2022) propõe JWSAN, um conjunto de dados de similaridade e associação de palavras em japonês. Além disso, o conjunto de dados fornece quatro tipos de pares de palavras: pares de palavras semanticamente semelhantes e que ocorrem frequentemente, pares de palavras semanticamente semelhantes e que ocorrem raramente, pares de palavras semanticamente diferentes e que ocorrem frequentemente e pares de palavras semanticamente diferentes e que ocorrem raramente.

JWSAN combina o uso de um dicionário de sinônimos para encontrar os pares de palavras que apresentam a mesma categoria semântica (espera-se que tenham alta similaridade) e os pares de palavras que apresentam diferentes categorias semânticas (espera-se que tenham baixa similaridade) e o uso de informação mútua pontual (PMI) que mede a tendência de co-ocorrência entre as palavras. Em outras palavras, o PMI ajuda a encontrar os pares de palavras que compartilham o mesmo contexto porque frequentemente co-ocorrem (espera-se que

tenham alta associação), e é improvável que os pares de palavras no mesmo contexto co-ocorram (espera-se que tenham baixa associação). (Vulić *et al.*, 2020) propõe um procedimento unificado para construção de conjuntos de dados para tarefas de similaridade de palavras em 12 idiomas, incluindo os idiomas de poucos recursos, como estoniano, finlandês, polonês, entre outros.

Em particular, nos concentramos em um problema do mundo real com um padrão objetivo para identificar boletins policiais semelhantes. Este trabalho também é o primeiro a comparar abordagens de similaridade semântica usando modelos pré-treinados ou de representação neural treinados com nossos dados no contexto de boletins policiais. Também investigamos diferentes representações de texto como entrada e outras métricas de similaridade. Além disso, estamos interessados em descobrir se a combinação de embeddings de sentença poderiam alavancar a busca por boletins policiais semelhantes. O Quadro 1 mostra o comparativo entre os trabalhos relacionados e o proposto.

Quadro 1 – Comparativo dos trabalhos relacionados e o proposto

Trabalhos	Idioma PT-BR	Treinamento	Similaridade entre Documentos	Combinação de Embeddings	Tarefa
RODRIGUES; MARCACINI, 2022	SIM	pré-treinado	SIM	SIM	Sentenças em português
DO; PHAM, 2021	NÃO	pré-treinado	SIM	SIM	Identificar entidades semelhantes
FIRMIANO; SILVA, 2021	SIM	pré-treinado e treinado	SIM	NÃO	Duplicação em boletins de ocorrência
IMTIAZ et al., 2020	NÃO	treinado	SIM	NÃO	Questões duplicadas
SHAHMIRZADI et al., 2019	NÃO	treinado	SIM	NÃO	Duplicação em patentes
NAILI et al., 2017	NÃO	treinado	SIM	NÃO	Segmentos de tópicos
INOHARA, K.; UTSUMI, 2022	NÃO	proposta de modelo (JWASAN)	SIM	NÃO	Semelhança entre palavras em japones
PROPOSTA	**SIM**	**pré-treinado e treinado**	**SIM**	**SIM**	**Duplicação em boletins de ocorrência**

Fonte: elaborado pelo autor (2023).

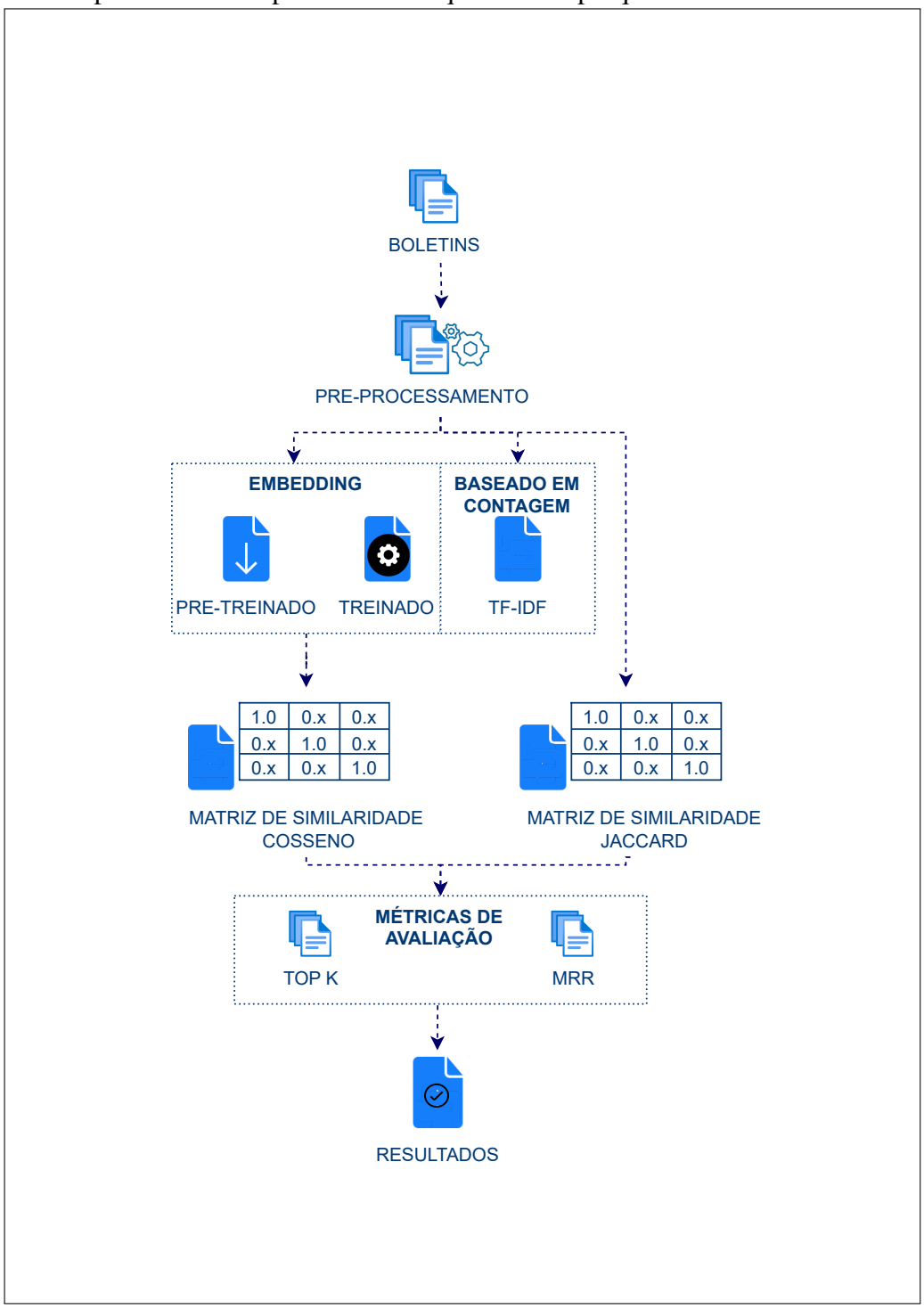
4 METODOLOGIA

Neste Capítulo, os dados e métodos usados em nosso problema serão discutidos. Dividimos nossa abordagem em duas etapas principais: (1) Gerar a matriz de similaridade para as diferentes representações dos boletins, tais como representações via modelos neurais e métodos de contagem de palavras ou *tokens*. Esta etapa refere-se ao pré-processamento dos dados, representação dos boletins, treinamento dos modelos e classificação dos K boletins mais semelhantes. A similaridade é calculada usando duas funções: Cosseno e similaridade de Jaccard. Porém, como explicamos, a similaridade de Jaccard é calculada mantendo as palavras na frase sem representá-las como vetores, mas como um conjunto de palavras. (2) Validar a precisão e o MRR (Mean Reciprocal Rank) de cada boletim policial do conjunto de dados de acordo com o seu boletim duplicado. Nesta etapa, a precisão do modelo é calculada validando se podemos recuperar o boletim policial mais semelhante dentro dos K principais boletins mais semelhantes indicadas por cada modelo. Também calculamos o MRR (Mean Reciprocal Rank), para avaliar a precisão dos modelos. A Figura 3 apresenta os passos que seguimos neste artigo.

Vale mencionar as razões para escolher e comparar as diferentes metodologias de representação de texto, além de resolver o problema de busca de similaridade de texto. Para abordar adequadamente os problemas de aprendizado de máquina, é essencial ter uma representação eficaz do boletim que ajude a máquina a entender o contexto. Pode ser útil oferecer resultados experimentais comparativos para os métodos, para que os pesquisadores possam determinar qual modelo de representação é mais apropriado para seu problema com base na análise comparativa.

No domínio de Processamento de Linguagem Natural (PLN), (Toshevska *et al.*, 2020; Boggust *et al.*, 2022) fornecem diferentes comparações entre embedding de sentenças para garantir a qualidade da representação de palavras antes do uso em uma tarefa de aprendizado de máquina. Os métodos de avaliação são classificados em duas categorias principais: intrínseco e extrínseco (Zhai *et al.*, 2016; Qiu *et al.*, 2018). Como já explicado anteriormente. Nossa avaliação é classificada como extrínseca e pode ser útil para os cientistas de dados na seleção do modelo de representação mais adequado, uma vez que é voltada para a tarefa de PLN, busca por similaridade.

Figura 3 – Etapas necessárias para avaliar as questões de pesquisa abordadas em nosso trabalho



Fonte: elaborado pelo autor (2021).

4.1 Conjunto de dados

A Secretaria de Segurança Pública do Ceará, Brasil, e a Secretaria de Segurança Pública do Pará, Brasil, forneceram os conjuntos de dados utilizados neste trabalho. O conjunto de dados é composto por dois *corpus*: um com 1.089 (onde, 1.065 boletins são similares,

contendo ao menos uma cópia para cada boletim de ocorrência) e outro com 30.011 boletins de ocorrência. Ambos os *corpus* correspondem a ocorrências eletrônicas e não eletrônicas entre 1º de janeiro de 2020 e 29 de março de 2020, relacionadas a diferentes tipos de crimes, como furto, homicídio, roubo, tentativa de homicídio e assédio moral.

O primeiro *corpus* com 1.089 boletins de ocorrência, possuí 1.065 boletins ao todo, em que cada boletim de ocorrência tem um duplicado neste mesmo conjunto. Os 24 boletins que não possuíam boletins mapeados como similares pela Secretaria de Segurança Pública foram descartados da validação, tendo em vista que não apresentavam nenhuma duplicidade no conjunto de dados. A Figura 4 demonstra uma parte do conjunto de dados com todas as informações disponibilizadas, como, relato do ocorrido, número do boletim, tipo de crime, endereço, data do registro, data do fato, entre outros.

Figura 4 – Conjunto de dados com os 1.089 boletins antes do pré-processamento e remoção dos boletins sem duplicidade.

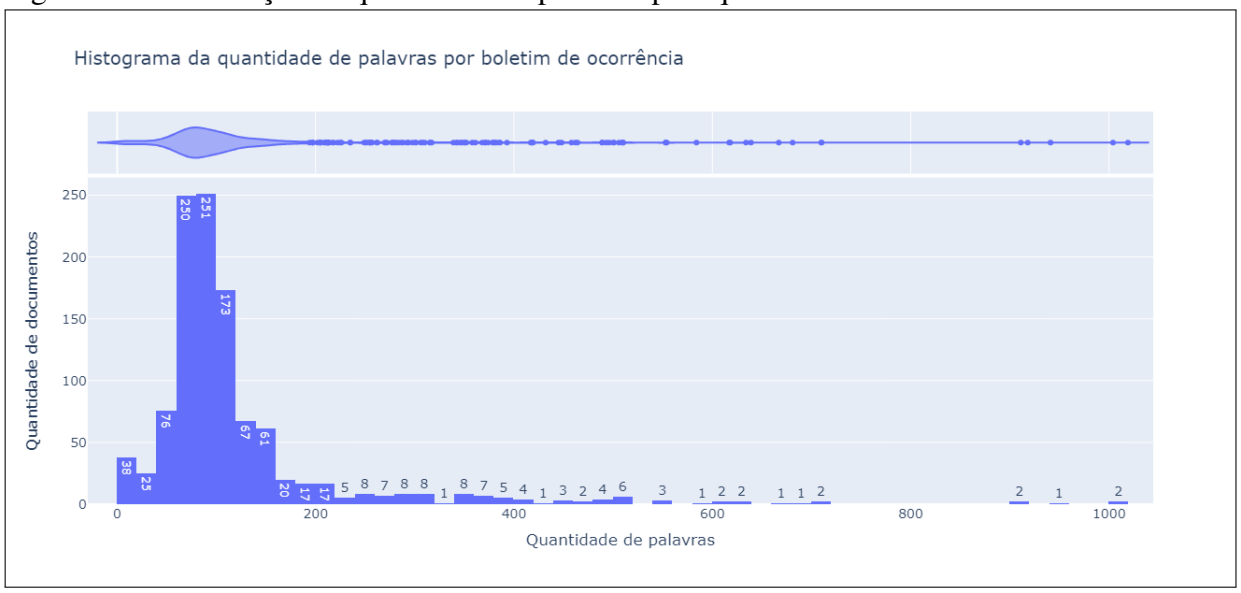
	RELATO	n_bop	n_tombo	tipo_tombo	unidade_origem	unidade_responsavel	data_registro	data_fato	registro	consolidado	fato_real	especificacao_crime	municipios	rua_fato	identificacao_do_fato
0	<p>O nacional relata QUE estava em casa QUE do...	00139/2020.100095-0	NAO INSTAURADO	NAO INSTAURADO	PORTO DE MOZ - DELEGACIA DE POLICIA - 11Âª RISP	PORTO DE MOZ - DELEGACIA DE POLICIA - 11Âª RISP	2020-03-25	2020-03-23	ROUBO AUMENTO DE PENA - EMPREGO DE ARMA	ROUBO	ROUBO	RESIDENCIA/CELULAR	PORTO DE MOZ	TRAVESSA SEM DENOMINACAO	TÃ_x008d_PICA > ROUBO > DOS CRIMES CONTRA O PA...
1	O relator acima qualificado comunica, através ...	277/2020.060163-2	NAO INSTAURADO	NAO INSTAURADO	DEL.VIRTUAL - 277	PORTO DE MOZ - UNIDADE POLICIAL 139	2020-03-25	2020-03-23	ART.157 - ROUBO	JA LANCADO	JA LANCADO	B.O.-00139/2020.100095-0	PORTO DE MOZ	RUA [redacted]	ROUBO IP: 170.[redacted]
2	O relator acima qualificado comunica, através ...	277/2020.002794-7	NAO INSTAURADO	NAO INSTAURADO	DEL.VIRTUAL - 277	ITUPIRANGA - UNIDADE POLICIAL 157	2020-01-06	2020-01-04	ART.157 - ROUBO	ROUBO	ROUBO	TRANSEUNTE/CELULAR	ITUPIRANGA	RUA [redacted]	ROUBO IP: 189.[redacted]
3	O relator acima qualificado comunica, através ...	277/2020.002762-1	NAO INSTAURADO	NAO INSTAURADO	DEL.VIRTUAL - 277	ITUPIRANGA - UNIDADE POLICIAL 157	2020-01-06	2020-01-04	ART.157 - ROUBO	JA LANCADO	JA LANCADO	B.O.-277/2020.002794-7	ITUPIRANGA	RUA [redacted]	ROUBO IP: 189.[redacted]
4	O relator acima qualificado comunica, através ...	277/2020.031592-2	NAO INSTAURADO	NAO INSTAURADO	DEL.VIRTUAL - 277	ANANINDEUA - SECCIONAL 28 - 2Âº RISP- 20Âª AISP	2020-02-11	2020-02-09	ART.157 - ROUBO	ROUBO	ROUBO	TRANSEUNTE	ANANINDEUA	RUA [redacted]	ROUBO IP: 131.[redacted]

Fonte: elaborado pelo autor (2023).

A Figura 5 apresenta algumas informações importantes que podem ser analisadas da seguinte forma:

- Distribuição de palavras por coluna: O número de palavras referente a cada coluna varia em aproximadamente 20 palavras, com a primeira coluna contendo 38 boletins (para até 20 palavras em cada boletim). É possível observar que a quantidade de boletins diminui conforme avançamos para as colunas subsequentes, ou com mais palavras.
- Extensão dos boletins: O menor boletim do conjunto de dados possui apenas 1 palavra, enquanto o maior boletim possui 1.019 palavras. Essa variação indica que os boletins têm tamanhos bastante distintos, mas grande parte está entre 70 a 120 palavras.
- Tamanho médio em termos de palavras: O tamanho médio dos boletins em termos de palavras é de 119,20. Isso sugere que, em média, os boletins são relativamente curtos, o

Figura 5 – Distribuição da quantidade de palavras pela quantidade de documentos.



Fonte: elaborado pelo autor (2023).

que dificulta o contexto.

– Tamanho médio em termos de caracteres: O tamanho médio dos boletins em termos de caracteres é de 597,41. Esse valor indica que, em média, os boletins têm uma baixa extensão em caracteres.
– Desvio padrão da quantidade de palavras: O desvio padrão da quantidade de palavras é de 111,03. Isso significa que a variação no tamanho das palavras entre os boletins é relativamente alta, indicando uma diversidade de comprimentos nos boletins.
– Desvio padrão do tamanho dos caracteres: O desvio padrão do tamanho dos caracteres é de 537,09. Esse valor mostra uma grande dispersão no tamanho dos boletins em termos de caracteres, ressaltando a diferença de comprimentos entre eles.
– Assimetria à direita do gráfico: O gráfico apresenta uma assimetria à direita, sugerindo que há uma maior concentração de boletins com menos palavras e um número menor de boletins com um grande número de palavras em suas descrições.

Essas informações são relevantes para entender a distribuição e a variabilidade dos boletins no conjunto de dados, bem como destacar a presença de boletins com baixo número de palavras, o que pode ser útil para análises posteriores.

4.2 Pré Processamento

O objetivo principal dessa etapa é remover informações irrelevantes e padronizar o texto de forma a garantir a consistência dos dados. O processo de remoção de caracteres

especiais é realizado para excluir elementos desnecessários, como códigos HTML e caracteres non-ASCII, que podem interferir na compreensão do texto. Além disso, os termos que contêm dígitos são eliminados, uma vez que esses elementos não são considerados relevantes para a análise de similaridade textual.

Outra tarefa importante é a alteração dos caracteres para minúsculas, o que torna o texto uniforme e evita que o modelo atribua pesos diferentes a palavras que aparecem em maiúsculas e minúsculas. Além disso, são removidas as palavras irrelevantes presentes nos documentos (stopwords), que são termos comuns na língua e que não agregam significado para a análise de similaridade.

O exemplo abaixo ilustra um boletim antes e depois de ser pré-processado e extraído do conjunto de dados contendo 1.065 boletins. Veja como o texto foi tratado durante o processo de pré-processamento:

- Antes: “O relator acima qualificado comunica, através da Delegacia Virtual, que no dia e hora acima mencionados foi vítima da seguinte ação criminosa: roubo, conforme o mesmo descreve abaixo: na noite do ultimo domingo, sair com a minha familia para lanchar, e na volta um carro branco que nao soubemos identificar parou ao nossos lado e fomos roubados, eles roubaram o celular e o RG da minha filha [NOME DA FILHA]. Quando perguntado se era possível identificar ou descrever os suspeitos, o relator assim escreveu: nao".

- Depois: "relator qualificado comunica através delegacia virtual dia hora mencionados vítima seguinte ação criminosa roubo descreve abaixo noite ultimo domingo sair minha familia lanchar volta carro branco nao soubemos identificar parou nossos lado roubados roubaram celular rg minha filha [NOME DA FILHA] perguntado possível identificar descrever suspeitos relator assim escreveu nao".

O pré-processamento de dados é fundamental para garantir que o modelo de representação neural possa trabalhar com um texto padronizado e sem informações irrelevantes que possam afetar o resultado final.

4.3 Representação das sentenças ou documentos

Podemos representar os boletins policiais com três soluções diferentes: através de um embedding pré-treinado, treinando um modelo de representação neural com nossos dados ou combinando diferentes embeddings para representar a sentença.

Neste trabalho, usamos como modelo de representação neural pré-treinado as seguin-

tes soluções: Flair, Universal Sentence Encoder (USE), Doc2Vec, Word2Vec, BERT, Sbert e GPT. Primeiro executamos os embeddings dos boletins e, em seguida, calculamos a média das representações de palavras para descrever a narrativa do boletim policial usando modelos de representação de palavras, como o Word2Vec.

Também investigamos o desempenho do Word2Vec, Doc2Vec e RoBERTa quando eles são treinados com o nosso conjunto de dados de boletins de ocorrência. Ambos os modelos foram treinados com dados da Secretaria de Segurança Pública do Ceará, Brasil, com cerca de 30 mil denúncias. Cada palavra recebeu sua representação inicial, que, durante o treinamento, foram realizados cálculos pelo modelo para obter uma melhor representação daquela palavra.

Para comparar o desempenho dos modelos para nossa tarefa e as abordagens lexicais, avaliamos o uso de um método baseado em contagem como TF-IDF, tal que cada palavra em um boletim é representada por seu TF-IDF. Na representação para similaridade Jaccard, usamos cada boletim policial representado por um conjunto de palavras.

4.4 Matriz de Similaridade

Depois de representar cada boletim policial como um vetor (usando um modelo de representação ou TF-IDF), podemos comparar os boletins policiais para encontrar os mais semelhantes. Para cada abordagem de representação investigada neste trabalho, geramos uma matriz NxN para classificar os K melhores boletins policiais mais semelhantes para cada boletim de ocorrência. Por exemplo, para o conjunto de dados com 1,065 boletins policiais representados com o Word2Vec, a comparação de similaridade de cosseno foi feita para todas os 1,065 boletins entre si, gerando assim uma matriz 1,065 x 1,065, contendo a porcentagem de similaridade.

Ignoramos a diagonal da matriz, pois a similaridade entre os mesmos documentos sempre seria de 100%. Claro, como dissemos antes, geramos uma matriz de similaridade de cosseno para cada representação, não apenas para Word2Vec, mas também para as abordagens Flair, Universal Sentence Encoder, Doc2Vec, Sbert, RoBERTa, BERT e TF-IDF. Também investigamos a manutenção das palavras no documento sem representá-las como vetores. Assim, geramos a matriz de similaridade de Jaccard. Nesse caso, a similaridade de Jaccard compara dois boletins policiais, verificando se elas têm palavras idênticas. A similaridade está em uma faixa de 0% a 100%. Quanto maior a porcentagem, mais semelhantes são os dois boletins policiais.

4.5 Ranking de Sentenças

Durante a etapa de classificação, são avaliados os principais top K (top 1, top 5 e top 10) e o MRR (*Mean Reciprocal Rank*). Para cada valor de K, é gerada uma lista dos documentos mais similares. Por exemplo, no caso do top 1, apenas o boletim mais similar ao texto analisado será listada (desconsiderando a diagonal da matriz de similaridade). Para o top 5, serão listadas os cinco boletins mais semelhantes, e para o top 10, os dez boletins mais similares.

Imaginemos que a lista de IDs de boletins policiais mais similares apresente o resultado correto, ou seja, que a lista contenha o boletim policial duplicado. Isso é considerado um acerto para a abordagem utilizada para representar os boletins (Flair, USE, Doc2Vec, Word2Vec, RoBERTa, BERT, Sbert, GPT ou a combinação desses modelos de representação, TF-IDF e sem representação vetorial, ou seja, mantendo as palavras nos documentos para calcular a matriz de similaridade de Jaccard).

4.6 Validação

Esse passo consiste em calcular a precisão de uma abordagem para encontrar os boletins mais semelhantes em um conjunto de dados. Para validar a acurácia dos modelos, foi utilizada a coluna de boletim duplicado no conjunto de dados que foi previamente mapeado, conferido e entregue pela Secretaria de Segurança Pública. A precisão é calculada como a porcentagem de acertos que a abordagem alcançou em relação a todo o conjunto de dados. Segue a formula utilizada para calcular a acurácia:

$$\text{Acurácia} = \frac{\text{Número de acertos}}{\text{Total de boletins}} \times 100\%$$

A Figura 6 representa como as informações do boletim e seu boletim duplicado foram representados para auxiliar na validação da assertividade dos modelos. As colunas "Nº_bo", "Nº_bo_similar"e "index_similar_bo"foram utilizadas para montar nossa validação. Abaixo segue a descrição de cada coluna:

- Nº_bo: Representa o número do boletim de ocorrência, que seria o seu identificador principal.
- Nº_bo_similar: Quando diferente de vazio, representa o número do boletim principal, ou seja, o primeiro boletim lançado sobre a ocorrência.
- index_similar_bo: Quando diferente de vazio, representa o identificador do boletim principal, utilizado para realizar as buscas no conjunto de dados.

Figura 6 – Representação dos identificadores do boletim e o seu identificador de duplicidade

	Nº_bo	Nº_bo_similar	index_similar_bo
0	0013920201000950		
1	27720200601632	0013920201000950	0
2	27720200027947		
3	27720200027621	27720200027947	2
4	27720200315922		
5	27720200315664	27720200315922	4
6	27720200315808	27720200315922	4
7	0029220201006716		
8	0029220201006701	0029220201006716	7
9	27720200371328		
10	27720200367300	27720200371328	9
11	27720200321450		

Fonte: elaborado pelo autor (2023).

Além disso, é calculado o MRR (*Mean Reciprocal Rank*) para cada método, que fornece uma média harmônica dos rankings dos documentos mais semelhantes encontradas pela abordagem. Esse processo ajuda a avaliar a eficácia das abordagens de busca de boletins mais semelhantes e determinar quais são mais precisas. Vale ressaltar que algumas outras métricas de avaliação poderiam ser utilizadas para validar a assertividade dos modelos, abordaremos algumas opções na seção de trabalhos futuros.

5 RESULTADOS

Neste Capítulo, conduzimos uma avaliação dos resultados obtidos através da utilização de modelos de representação de sentenças pré-treinados. Também examinamos o impacto da utilização de modelos treinados em um novo conjunto de dados de boletins policiais, na identificação de documentos semelhantes, incluindo boletins com sobreposição de palavras e registros de mesmo *modus operandi*. Além disso, analisamos o desempenho de métodos de representação de texto baseados em contagem, como o TF-IDF, e investigamos se a combinação de múltiplas representações de embedding de sentenças poderia aprimorar a busca por boletins policiais similares.

As perguntas de pesquisa apresentadas na introdução desta pesquisa guiaram a análise realizada neste Capítulo, a qual visa aprofundar nosso entendimento sobre os métodos de identificação de duplicatas de boletins policiais. Compreender o desempenho de diferentes modelos e técnicas de representação de texto é fundamental para o desenvolvimento de soluções mais eficazes e precisas na identificação de boletins de ocorrência duplicados.

Antes de adentrarmos nos resultados obtidos em cada uma das perguntas de pesquisa, é importante que apresentemos alguns exemplos do conjunto de dados utilizados. Como mencionado anteriormente, o nosso conjunto de dados é composto por 1.065 boletins policiais, os quais possuem, além do boletim original, a indicação de qual outro boletim é uma duplicação.

O Quadro 2 abaixo apresenta dois exemplos concretos para ilustrar a estrutura do nosso conjunto de dados. É possível observar que os boletins duplicados apresentam pequenas variações entre si, as quais podem ser oriundas de diferenças nas palavras utilizadas pelos policiais que os redigiram, por exemplo. Essa estrutura de dados é de grande relevância para a realização da presente pesquisa, pois nos permite analisar a presença de duplicações de boletins policiais e entender suas causas e impactos.

Compreender as causas da duplicação de boletins policiais é importante para evitar retrabalho e atrasos desnecessários no processamento desses documentos. Uma das principais razões é que a mesma ocorrência pode ser registrada por diferentes vítimas, cada uma usando suas próprias palavras para descrever o evento. Isso é especialmente comum em casos de crimes que afetam várias pessoas, como assaltos em locais públicos.

Outra causa comum de duplicação de boletins policiais é a falha do sistema ou do usuário ao registrar o boletim eletrônico. Por exemplo, um usuário pode ter registrado um boletim, mas devido a um erro no sistema, não foi salvo corretamente. Posteriormente, o

Quadro 2 – Exemplos de boletins de ocorrência e o boletim duplicado.

Boletim de Ocorrência	Boletim de Ocorrência Duplicado
O relator acima qualificado comunica, por meio da Delegacia Virtual, que no referido dia e horário foi vítima da seguinte ação penal: furto, conforme abaixo descrito: Dois elementos em uma motocicleta POP de cor preta, com uma faca em punho , chegou fazendo ameaças. Gritando: "me dá, me dá o celular". Eles levaram um celular da marca Samsung, A205GT, com o número de série: 357621102968577. Questionado se era possível identificar ou caracterizar os suspeitos, o relator escreveu: “Não, não foi possível identificá-los, pois usavam capacetes."	O relator acima qualificado comunica através da Delegacia Virtual, que no dia e horário supracitados foi vítima da seguinte ação penal: furto, conforme abaixo descrito: dois elementos em uma motocicleta POP preta, com ameaça de faca, chegaram falando passa o celular, passa o celular. Questionado se era possível identificar ou caracterizar os suspeitos, o relator escreveu: não, usavam capacete
O relator acima qualificado comunica, por meio da Delegacia Virtual, que no referido dia e horário foi vítima da seguinte ação penal: furto, conforme abaixo descrito: Eu havia acabado de sair do trabalho e estava no ponto de ônibus, na Avenida Nazaré, entre a rua Rui Barbosa e a rua Quintino, quando 4 pessoas armadas desceram de um carro e assaltaram a mim e outras pessoas que ali estavam, levaram minha bolsa contendo meu celular, cartão de banco e uma quantia de aproximadamente 60 reais. Questionado se era possível identificar ou caracterizar os suspeitos, o relator escreveu: eram 5 pessoas, uma ficou dentro do carro esperando na Rua Rui Barbosa, as outras 4 estavam bem vestidas, todos de bermuda e camisa normal, um deles tinha o cabelo tingido de loiro.	O relator acima qualificado comunica, por meio da Delegacia Virtual, que no dia e horário supracitados foi vítima da seguinte ação penal: furto, conforme abaixo descrito: Tinha acabado de sair do trabalho e estava no ponto de ônibus da rua Nazaré Avenida, entre a Rua Rui Barbosa e a Rua Quintino Bocaiuva, quando quatro homens armados desceram de um carro e assaltaram eu e outras pessoas que estavam ali, gritaram e empurraram todos apontando suas armas, levaram minha bolsa contendo meu celular, cartão do banco e valor aproximado de R$ 60,00. Questionado se era possível identificar ou caracterizar os suspeitos, o relator escreveu: Eram cinco pessoas, uma ficou no carro esperando as demais na Rua Rui Barbosa, as outras quatro estavam todas bem vestidas, de bermuda e camisa de manga, um deles pintou o cabelo de loiro, fugiram de carro pela Rua Rui Barbosa

Fonte: elaborado pelo autor (2023)

usuário pode tentar registrar o mesmo boletim novamente, sem perceber que já havia registrado anteriormente. Essas falhas podem ocorrer em qualquer sistema eletrônico e podem ser difíceis de detectar manualmente.

Além dessas causas comuns, outras possibilidades de duplicação de boletins policiais incluem erros de registro devido à falta de informações precisas ou à falta de uma identificação clara da ocorrência. Também é possível que haja duplicações por engano, fraude, necessidade ou má intenção. Portanto, é importante que as autoridades policiais implementem medidas de prevenção e detecção de duplicatas, como o uso de sistemas eletrônicos que possam identificar automaticamente boletins duplicados e a realização de verificações regulares de qualidade para garantir que os dados registrados sejam precisos e confiáveis. Vale ressaltar que os motivos aqui apresentados também foram informados pela Polícia Civil do estado do Pará, concedente dos

dados.

Estudo sobre os resultados da QP1. A primeira pergunta de pesquisa investiga o melhor método de representação para capturar a similaridade sintática entre os boletins policiais. O *corpus* empregado para explorar a problemática em questão consiste em um conjunto de dados composto por 1.065 boletins policiais. Esse conjunto de dados foi fornecido com boletins já avaliados, ou seja, temos o índice do boletim policial duplicado identificado no conjunto de dados. Quando analisamos esse conjunto de dados, observamos que o boletim policial e sua duplicata apresentam palavras sobrepostas.

Avaliamos como os modelos de representação pré-treinados classificaram os boletins duplicadas e os métodos baseados em contagem usando a representação TF-IDF. Para tais abordagens, comparamos os boletins por meio da similaridade cosseno. Também investigamos o uso de outra estratégia básica: a similaridade Jaccard, para isso mantivemos as palavras nas frases sem representá-las como vetores numéricos. Os modelos pré-treinados avaliados foram o USE, Word2Vec, Doc2Vec, BERT, SBert, Flair e GPT. A Tabela 1 mostra os resultados de precisão e MRR obtidos pelas abordagens.

Tabela 1 – Resultados dos modelos na avaliação da **QP1**.

Modelo	Configuração		Assertividade			
	Pre-embedding	Dim.	Top 1	Top 5	Top 10	MRR
USE	`universal-sentence-encoder-multilingual/3`	512	70.47%	80.62%	82.61%	75.19
BERT	`quora-distilbert-multilingual`	768	65.58	74.46	76.99	69.85
Word2Vec	`cbow_s50`	50	64.86	73.19	75.91	69.06
SBert	`distiluse-base-multilingual-cased-v1`	512	63.04	73.37	76.09	67.95
Flair	`bert-base-nli-mean-tokens`	768	51.45	60.14	61.96	55.75
Doc2Vec	`pt_wiki_corpus`	150	35.87	47.46	51.27	43.39
GPT	`text-embedding-ada-002`	1536	31.52	33.51	34.06	32.81
TF-IDF	N/A	N/A	**77.38**	**86.05**	**87.68**	**81.58**
Jaccard	N/A	N/A	75.18	83.15	85.51	79.14

Fonte: elaborado pelo autor (2023).

A representação do boletim usando o modelo USE se sobressaiu em relação aos outros modelos de representação. O modelo USE é baseado em *Transformer Encoder* e apresenta maior precisão do que a versão com um modelo de rede profunda média (DAN) como codificador. O resultado que alcançamos é quase a mesma precisão relatada no artigo (Cer *et al.*, 2018) para o STS Bench (Banco de similaridade textual semântica). Ele realiza o embedding dos boletins em um vetor com 512 dimensões, o que possibilita uma representação semântica mais precisa. O modelo utilizado foi o multilingue, o qual representa sentenças de diferentes idiomas, além do

inglês e do português.

O segundo melhor resultado, em termos de modelos de embedings, foi obtido pela representação de outro modelo baseado em *Transformer*. BERT obteve o segundo melhor resultado para os modelos que capturam a síntaxe e a semântica na representação. O pré-treinado BERT obtido em (Souza *et al.*, 2020) superou ligeiramente o pré-treinado Word2Vec obtido em (Hartmann *et al.*, 2017). Já o modelo SBert multilingual com 512 dimensões obteve um resultado bem próximo ao Word2Vec, chegando a superá-lo no top 5 e top 10. O modelo Flair utilizando a configuração de 768 dimensões não obteve resultados significativamente melhores que os outros modelos, ficando abaixo dos demais.

Para o modelo pré-treinado Doc2Vec, utilizou-se o conjunto de dados da Wikipedia em português. Dessa forma, os dados apresentam uma ampla variedade de frases utilizadas em diferentes contextos, o que pode ter afetado negativamente o desempenho do modelo na tarefa de identificação de boletins similares. Como resultado, observou-se uma baixa precisão na identificação dos boletins mais similares, o que pode comprometer a sua confiabilidade em aplicações práticas.

O GPT utilizando o modelo *text-embedding-ada-002* (Neelakantan Tao Xu, 2022), obteve um baixo desempenho no experimento, com um nível de assertividade abaixo dos demais modelos. O GPT Embeddings Ada-002 é um modelo de linguagem treinado em uma tarefa de previsão de palavras seguintes (ou seja, tarefa de autoregressão), enquanto o Universal Sentence Encoder, BERT e outros, são treinados em tarefas de codificação de sentenças, como previsão de sentenças seguintes ou classificação de sentenças. Isso pode resultar em diferenças sutis na forma como os modelos aprendem a representação das sentenças, levando a resultados diferentes na busca de similaridade. Além disso, o GPT Embeddings Ada-002 é treinado de forma unidirecional, ou seja, ele prevê palavras apenas com base no contexto anterior, enquanto o BERT é treinado de forma bidirecional, considerando o contexto anterior e posterior de cada palavra. Isso permite ao BERT capturar informações contextuais de maneira mais abrangente, o que pode ser vantajoso na busca de similaridade entre boletins.

O TF-IDF foi calculado usando o *TfidfVectorizer* [1] para representar os boletins. O primeiro passo foi separar cada palavra entre as sentenças e depois criar um dicionário com cada palavra. As *stopwords* em português foram removidas para evitar variação na similaridade entre diferentes sentenças. Depois disso, calculamos a frequência dos termos para obter o número de

[1] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

vezes que cada palavra apareceu no documento. Em seguida, calculamos a frequência inversa para obter as expressões mais raras, ou seja, com menos recorrência. Quando analisamos os resultados, a representação do TF-IDF obteve bons resultados. Um indicativo é por conta do tamanho em número de palavras dos textos, a maioria são textos curtos.

O que podemos aprender com esse experimento é que quando os boletins apresentam bastante sobreposição de palavras, a representação por meio da média dos embeddings de palavras ou do embedding de sentença pode ser pior do que comparar as palavras de maneira bruta (observe que TF-IDF e Jaccard superam os outros), especialmente em textos curtos, mesmo que esses embeddings tenham sido supostamente treinados para capturar semelhança semântica e sintática. Ainda há espaço para melhoria nos embeddings pré-treinados.

Estudo sobre os resultados da QP2. Nesta pergunta de pesquisa, investigamos como os modelos pré-treinados se comportariam ao usar palavras diferentes no contexto de crimes semelhantes, ou seja, a assertividade na similaridade semântica entre os boletins. A Tabela 2 mostra os resultados obtidos por esses modelos. A partir do conjunto de dados com 30 mil registros policiais, separamos manualmente 50 registros policiais (25 pares com o mesmo *modus operandi*, sem demasiada sobreposição de palavras). O conjunto de dados consiste em diferentes crimes, como roubo, furto, assédio sexual e outros.

Tabela 2 – Resultados dos modelos na avaliação da **QP2**.

Modelo	Configuração		Assertividade			
	Pre-embedding	Dim.	Top 1	Top 5	Top 10	MRR
SBert	`distiluse-base-multilingual-cased-v1`	512	46.00	**84.00**	**100.00**	62.76
USE	`universal-sentence-encoder-multilingual/3`	512	46.00	84.00	90.00	**63.08**
BERT	`quora-distilbert-multilingual`	768	36.00	74.00	90.00	53.52
Word2Vec	`cbow_s50`	50	20.00	54.00	68.00	35.16
Doc2Vec	`pt_wiki_corpus`	150	28.00	64.00	80.00	43.79
Flair	`bert-base-nli-mean-tokens`	768	20.00	38.00	50.00	31.15
TF-IDF	N/A	N/A	**48.00**	68.00	88.00	61.67
Jaccard	N/A	N/A	44.00	68.00	80.00	56.63

Fonte: elaborado pelo autor (2023).

O modelo SBert obteve o melhor resultado nos top 5 e top 10 boletins mais semelhantes, atingindo 100% de assertividade na detecção de boletins semelhantes no top 10. De fato, ele também supera a forma bruta para comparar os boletins via TF-IDF e a similaridade de Jaccard. O modelo USE ficou bem próximo dos resultados do SBERT, conseguindo obter um resultado mais preciso no MRR. Para entender esses resultados, considerando a similaridade entre 0 e 100, onde 0 representa nenhuma similaridade, e 100 representa a similaridade máxima.

A Tabela 3 mostra a similaridade média entre um boletim policial e o indicado como o mais semelhante pelos modelos avaliados, além da média de similaridade do texto para o mais similar de acordo com o *ground truth* (realmente similar).

Tabela 3 – Resultado da média de similaridade entre o boletim mais similar para o modelo, comparado ao realmente similar.

Modelo	Configuração		Média de Similaridade	
	Pre-embedding	Dim.	Similar - Modelo	Ground truth
USE	`universal-sentence-encoder-multilingual/3`	512	64.41	60.95
BERT	`quora-distilbert-multilingual`	768	91.56	**89.82**
Word2Vec	`cbow_s50`	50	82.45	73.58
Doc2Vec	`pt_wiki_corpus`	150	60.48	43.79
Flair	`bert-base-nli-mean-tokens`	768	**91.71**	84.52
SBert	`distiluse-base-multilingual-cased-v1`	512	63.21	59.50
TF-IDF	N/A	N/A	29.50	27.40
Jaccard	N/A	N/A	20.24	18.86

Fonte: elaborado pelo autor (2023).

Observa-se que, dentre todos os modelos e métodos avaliados, o TF-IDF e a similaridade de Jaccard apresentaram, em média, os menores valores de similaridade. Isso indica que, apesar dessas abordagens terem se destacado em relação a outros modelos de representação de texto, como o Word2Vec e o Doc2Vec, sua confiabilidade é baixa. Em outras palavras, a similaridade entre um boletim policial e seu boletim mais semelhante é extremamente baixa quando utilizamos esses métodos, uma vez que eles medem a similaridade com base nas palavras presentes em ambos os documentos. Em contraste, os métodos de representação não dependem da presença exata das palavras nos documentos, mas sim da similaridade no espaço vetorial do embedding. Portanto, o TF-IDF e a similaridade de Jaccard seriam métodos mais fracos na busca por boletins policiais com o mesmo *modus operandi*.

Estudo sobre os resultados da QP3. A terceira pergunta de pesquisa investiga se os modelos de embedding de sentenças são mais eficazes na representação de boletins de ocorrência quando treinados com um conjunto de dados de boletins policiais. Usando uma abordagem não supervisionada, este estudo treinou os modelos RoBERTa (uma melhoria do modelo BERT), Doc2Vec e Word2Vec. Os dados para treinamento são os 30.000 boletins policiais fornecidos pela Secretaria de Segurança Pública do Pará (utilizamos boletim-pará como referência na Tabela 4), Brasil. Para investigar a **QP3**, usamos o conjunto de dados com 1.065 boletins policiais e sua duplicata (o mesmo usado para a **QP1**) representados pela Tabela 4.

O modelo treinado RoBERTa obteve uma maior assertividade na identificação de

Tabela 4 – Resultados dos modelos avaliados na **QP3**.

Modelo	Treinamento	Assertividade Dim.	Top 1	Top 5	Top 10	MRR
RoBERTa - treinado	boletim-pará	768	**68.66**	**77.17**	**79.89**	**72.75**
BERT	`quora-distilbert-multilingual`	768	65.58	74.46	76.99	69.85
Word2Vec - treinado	boletim-pará	50	65.22	73.73	75.36	69.36
Word2Vec	`cbow_s50`	50	64.86	73.19	75.91	69.06
Doc2Vec - treinado	boletim-pará	150	37.32	49.46	55.62	43.37
Doc2Vec	`pt_wiki_corpus`	150	35.87	47.46	51.27	43.39
TF-IDF	N/A	N/A	**77.38**	**86.05**	**87.68**	**81.58**
Jaccard	N/A	N/A	75.18	83.15	85.51	79.14

Fonte: elaborado pelo autor (2023).

documentos semelhantes. Treinamos o RoBERTa por 13 épocas quando a diferença entre a *função de loss* da época atual e a anterior não era mais significativa do que 0,01. O modelo RoBERTa gera embeddings de dimensionalidade 768. A representação dos boletins de ocorrência para o Word2Vec foi feita com vetores de 50 dimensões, treinados por 100 épocas. Para o modelo Doc2Vec, as representações foram geradas com vetores de 150 dimensões, treinados por 100 épocas, com uma janela de treinamento de tamanho 10.

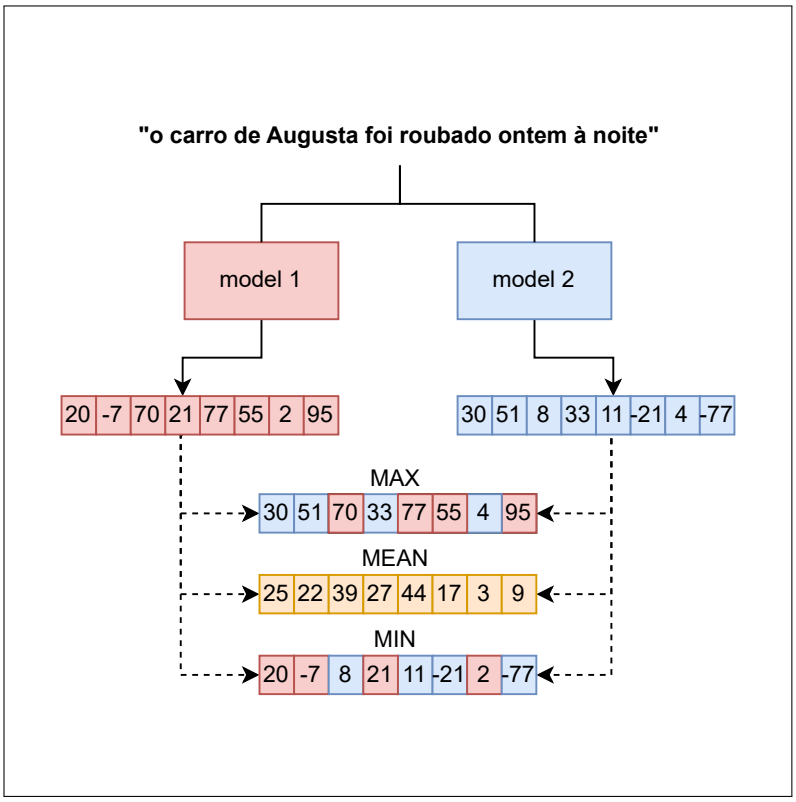
Observe que as três abordagens treinadas com os 30.000 boletins policiais (**Word2Vec - treinado, RoBERTa - treinado** e **Doc2Vec - treinado**) superaram ligeiramente os modelos correspondentes pré-treinados com um conjunto de dados de contexto diferente. Como já esperado, é preferível usar modelos de representação treinados com dados do mesmo contexto. Isso é esperado porque os embeddings de sentenças treinados e usados com dados do mesmo contexto devem manter a proximidade entre vetores semelhantes, que representam relações sintáticas, semânticas ou morfológicas. Por exemplo, palavras com significados diferentes devem ser colocadas em pontos distantes umas das outras, enquanto palavras usadas no mesmo contexto são colocadas em pontos próximos, pois geralmente aparecem em frases semelhantes. Os resultados são satisfatórios, considerando que as versões pré-treinadas do Word2Vec, BERT e Doc2Vec foram treinadas com um corpus muito maior, contendo 1.395.926.282 tokens, enquanto nossos 30.000 boletins policiais são significativamente menores, com apenas 2.579.815 tokens.

É importante observar que, no ambiente atual, caracterizado por boletins mais concisos, os métodos TF-IDF e Jaccard demonstraram alcançar resultados superiores em comparação aos modelos de aprendizado de máquina treinados e suas variantes pré-treinadas. Isso nos mostra que para trabalhos mais simples, faz-se viável o uso desses modelos. Entretanto, como discutido na questão de pesquisa anterior (**QP2**), a confiabilidade desses resultados é questionável, o que

nos leva a descartar a aplicação desses modelos em cenários reais ou em boletins mais extensos, nos quais as descrições dos eventos sejam mais detalhadas e específicas.

Estudo sobre os resultados da QP4. Esta questão de pesquisa investiga se a combinação de embeddings pode melhorar a busca por boletins policiais duplicados. Foram utilizados os dois melhores modelos de embedding dos resultados relatados na **QP1**. Além disso, também foi investigado se a combinação desses dois embeddings pode superar os modelos de embedding individuais na recuperação de boletins policiais com o mesmo *modus operandi*.

Figura 7 – Representação da etapa de obtenção dos valores MAX, MEAN, MIN da combinação de embeddings de dois modelos.



Fonte: elaborado pelo autor (2023).

O modelo Universal Sentence Encoder (USE) foi selecionado como base para nossa análise, uma vez que ele demonstrou ser o mais eficaz na identificação de sentenças similares. Com o objetivo de comparar a efetividade da combinação de embeddings, selecionamos os dois melhores modelos: RoBERTa e Word2Vec, ambos treinados com 30 mil boletins policiais. Como o modelo USE gera embeddings com 512 dimensões, foi necessário ajustar as configurações de treinamento do modelo Word2Vec para que ele também gerasse embeddings com 512 dimensões. O mesmo processo foi aplicado ao modelo RoBERTa, que inicialmente gerava embeddings de 768 dimensões. Para garantir que todos os boletins fossem representados na mesma dimensão,

optamos por utilizar o modelo *bert-base-multilingual-cased* com 512 dimensões.

A assertividade dos modelos USE + RoBERTa (USERoBERTa) e USE + Word2Vec (USEW2V) foi comparada para dois conjuntos de dados: os 1.065 boletins utilizados na **QP1** e os 50 boletins utilizados na **QP2**. A Figura 7 exemplifica o processo realizado entre os modelos abordados nessa questão de pesquisa para obter os valores MAX, MEAN, MIN. Temos a frase: "o carro de Augusta foi roubado ontem à noite", onde, cada modelo irá gerar um vetor numérico contendo a representação dessa frase. Vale ressaltar que, para essa abordagem, ambas representações devem conter o mesmo tamanho vetorial. Para cada índice, será comparado o máximo, mínimo e a média entre os dois vetores.

Tabela 5 – Resultados da combinação de embeddings dos modelos USE + RoBERTa e USE + Word2Vec para avaliação de **QP4**.

Modelo	Assertividade				
	Dim.	Top 1	Top 5	Top 10	MRR
USE	512	**70.47**	**80.62**	82.61	**75.19**
USERoBERTa MEAN	512	70.47	80.25	**82.79**	75.11
USERoBERTa MAX	512	70.47	79.17	81.70	74.77
USERoBERTa MIN	512	69.57	78.80	82.07	74.33
USEW2V MEAN	512	69.93	80.25	82.07	75.05
USEW2V MAX	512	69.20	79.53	81.52	74.34
USEW2V MIN	512	68.66	78.99	81.16	74.32

Fonte: elaborado pelo autor (2023).

Na Tabela 5, é possível analisar os resultados obtidos através das combinações entre os melhores modelos de embedding, chamados de: USERoBERTa e USEW2V, variando entre (MAX, MEAN, MIN).

Para os 1.065 boletins foram muito similares aos resultados obtidos usando apenas o modelo USE. O modelo USERoBERTa MEAN na métrica Top 10 apresentou uma pequena vantagem de 0,18 em relação ao modelo USE, porém, nos outros rankings permaneceu igual ou inferior. Como os 1.065 boletins apresentam palavras sobrepostas, não conseguimos obter resultados significativamente melhores combinando os dois modelos para representar os documentos com esse conjunto de dados.

Podemos observar que na Tabela 6, ao realizar a mesma comparação com os 50 boletins selecionados por tipo de crime e *modus operandi*, os modelos USERoBERTa e USEW2V (MAX, MEAN, MIN) superaram os modelos individuais. A combinação dos modelos superou o modelo USE em quase todas as comparações (Top 1, Top 5 e MRR), houve um aumento considerável na indicação do documento mais similar, exceto pelo top 10 das representações

usando USEW2V. Em síntese, a combinação de embeddings de modelos revelou-se tão eficaz quanto ou até superior à utilização dos modelos isoladamente para a nossa questão de pesquisa.

Tabela 6 – Resultados obtidos da combinação de embeddings dos modelos USE + RoBERTa e USE + Word2Vec em 50 boletins com *modus operandi* semelhantes na avaliação da **QP4**.

Modelo	Dim.	Assertividade Top 1	Top 5	Top 10	MRR
USE	512	46.00	84.00	**90.00**	63.08
USERoBERTa MEAN	512	48.00	84.00	**90.00**	64.68
USERoBERTa MAX	512	52.00	84.00	**90.00**	64.71
USERoBERTa MIN	512	48.00	86.00	**90.00**	64.79
USEW2V MEAN	512	54.00	**88.00**	88.00	68.78
USEW2V MAX	512	**62.00**	84.00	88.00	**70.89**
USEW2V MIN	512	52.00	**88.00**	88.00	67.04

Fonte: elaborado pelo autor (2023).

Discussão. A Tabela 7 fornece alguns exemplos de boletins policiais de nosso conjunto de dados com o mesmo *modus operandi*, além de qual documento o melhor modelo de embedding (de nossos experimentos, o modelo USE) indica como a mais similar (coluna chamada ID-USE) e a verdadeira (coluna chamada Real ID). Vale ressaltar que por motivos de confidencialidade os nomes e algumas informações sensíveis foram retiradas do texto.

As duas primeiras linhas ilustram um cenário no qual o modelo USE apresentou equívocos. Os documentos com ID 1 e 4 se referem a um evento de roubo, com uma mulher como vítima e o mesmo objeto roubado (celular da vítima). No entanto, os documentos com ID 1 e 2 são mais similares em termos de *modus operandi*. Ambas descrevem uma mulher como vítima; o objeto roubado é o mesmo (o cartão SIM) e há uma pessoa como ladrão portando uma faca em ambos os casos.

As duas últimas linhas da Tabela 7 mostram um caso em que o modelo USE encontrou corretamente os boletins mais similares com ID 3 e 4. Ambos os boletins policiais seguem o mesmo *modus operandis*, uma vez que ambos descrevem um roubo com uma arma de fogo por dois ladrões. Esses ladrões levaram os mesmos pertences pessoais da vítima, como celular, dinheiro e documentos de identificação. Nosso objetivo principal com esses exemplos é mostrar que pode ser desafiador para o modelo de representação indicar corretamente o documento mais similar em termos de *modus operandis*. Com base nos resultados obtidos, conclui-se que o modelo USE apresenta um desempenho confiável para auxiliar os policiais na busca por boletins policiais semelhantes, além de proporcionar uma compreensão mais profunda sobre a maneira

Tabela 7 – Comparação de similaridade entre boletins policiais com o mesmo *modus operandi*.

ID	Boletim	USE ID	ID Real
1	O relator acima qualificado comparece a esta seção para registrar que, na data, horário e local mencionados acima, sua filha menor de idade, [NOME], teria sido vítima de uma agressão. Ela estava caminhando em uma via pública quando foi abordada por um homem de altura mediana, gordo, cor de pele morena, que utilizava uma motocicleta preta, modelo Honda Fan, e portava uma faca. Sob forte ameaça, ele subtraiu o dispositivo de telefone celular da vítima, modelo Motorola Moto G4 Plus, de cor preta e capa vermelha, contendo dois cartões SIM das operadoras Tim e Claro. O registro é feito para fins apropriados.	4	2
2	O relator acima qualificado comparece a esta seção para registrar que, na data, horário e local mencionados acima, a menor [NOME] teria sido abordada por um homem com uma faca, que, sob forte ameaça, subtraiu os seguintes objetos que estariam com a menor: um dispositivo de telefone celular modelo Motorola, de cor preta, contendo dois chips, das operadoras Claro e Tim, cadernos, caneta e livros que estariam dentro da mochila roubada pelo autor do fato. O registro é feito para fins apropriados.	3	1
3	O relator acima qualificado comparece a esta delegacia de polícia na data, horário e local mencionados acima, onde teria sido vítima de roubo. Ele estava caminhando em uma via pública quando foi abordado por dois indivíduos de gêneros masculino e feminino, que utilizavam uma motocicleta colorida predominantemente vermelha. Sob forte ameaça, foi informado de que havia uma arma de fogo e, em seguida, os criminosos subtraíram os seguintes objetos do relator: seu RG, CPF, uma mochila contendo roupas, uma quantia em dinheiro de R$15,00, além de um dispositivo de telefone celular modelo Samsung, de cor cinza, contendo um chip da operadora Tim. Os criminosos fugiram imediatamente após o roubo. O registro é feito para fins apropriados.	4	4
4	O relator acima qualificado comparece a esta delegacia de polícia para registrar que, na data, horário e local mencionados acima, foi vítima de um assalto. Ela alega ter sido abordada por dois homens em uma via pública, que utilizavam uma motocicleta de cor predominante preta e o passageiro portava uma arma de fogo. Sob grave ameaça, os criminosos subtraíram a bolsa da relatora, contendo seu RG, CPF, uma quantia em dinheiro de R$80,00 e um celular modelo Samsung, de cor prata, com dois cartões SIM das operadoras Tim e Vivo. O registro é feito para fins apropriados.	3	3

Fonte: elaborado pelo autor (2023).

como os criminosos agem.

6 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, o objetivo principal foi abordar o desafio de encontrar o boletim policial mais semelhante em um banco de dados a partir de um boletim fornecido. Para atingir este objetivo, foi necessário considerar não apenas a similaridade lexical, mas também a semântica dos dados de texto. Além disso, em comparação com o trabalho anterior, conseguimos incluir uma variedade maior de modelos e aplicar mais combinações, auxiliando-nos na busca de similaridade em boletins de ocorrência.

Podemos responder ao decorrer deste trabalho algumas questões de pesquisa importantes, como qual modelo pré-treinado apresenta melhor resultado ao problema de identificação de boletins de ocorrência similares. Para isso, foram avaliados diversos modelos experimentais, sendo o USE, um modelo baseado em *Transformer* e multilíngue, apresentou o melhor desempenho. Apesar de o método não refinado de comparar os boletins por meio do TF-IDF e da similaridade de Jaccard ter obtido resultados similares aos do modelo USE, a confiabilidade desses métodos é limitada, uma vez que exibiram menor grau de semelhança entre o boletim em análise e o boletim mais semelhante indicada pelos referidos modelos.

Também avaliamos a precisão dos modelos em relação aos boletins policiais semanticamente semelhantes, que foram separados com base no *modus operandi* de 50 boletins policiais distintos. O desempenho do Universal Sentence Encoder foi novamente superior aos demais modelos, tanto em termos de top-K (K igual a 1, 5 e 10) quanto de MRR. Isso indica que o USE é capaz de capturar as nuances semânticas e sintáticas dos boletins policiais, e é uma opção confiável para identificar boletins de ocorrência similares.

Após realizarmos uma análise detalhada, chegamos à conclusão de que utilizar modelos de representação treinados com dados do mesmo contexto é a abordagem mais vantajosa. Esta conclusão foi embasada na observação de que os embeddings de sentença gerados por esses modelos mantêm a proximidade entre vetores semelhantes, o que permite representar relações sintáticas, semânticas ou morfológicas de forma mais precisa. Consequentemente, essa abordagem demonstrou ser mais eficaz na tarefa de busca de similaridade em boletins policiais, tornando-se uma alternativa promissora para a análise de boletins policiais em geral.

Este trabalho também investigou a abordagem de combinar diferentes embeddings usando diferentes funções de agregação, como máximo, mínimo e média. Em alguns casos, essa técnica foi mais eficaz na representação da narrativa e na identificação de boletins policiais semelhantes. Isso sugere que a combinação de diferentes abordagens de embeddings podem

levar a uma representação mais completa e precisa dos dados, resultando em melhores resultados na busca por boletins policiais semelhantes. Com os resultados alcançados, podemos concluir que este trabalho poderia ser aplicado no mundo real, auxiliando a polícia na detecção de boletins duplicados.

Para trabalhos futuros, é sugerido que essas abordagens sejam avaliadas em diferentes conjuntos de dados policiais e que os embeddings sejam investigados em diferentes bases de dados, de outros contextos. Outra sugestão seria utilizar outras formas de combinar embeddings, como a concatenação de duas representações obtidas por modelos diferentes, formando uma única. O uso de Contrastive Learning Embeddings, poderia ser utilizada para detecção de boletins similares, com a capacidade de aprender sem supervisão explícita, ou outros conceitos como *one-shot* e *zero-shot*.

Uma outra sugestão seria treinar os modelos em um conjunto de dados com descrições mais detalhadas dos crimes, documentos de processos criminais ou textos mais longos. Poderíamos concatenar as informações de vítima, local, data do ocorrido e outras informações para termos representações mais precisas. Notamos que alguns boletins careciam de detalhes em sua descrição, contendo apenas um breve resumo do ocorrido informando o tipo de crime (furto, roubo, latrocínio, etc.) mas sem o detalhamento do *modus operandi* do fato. O que levou a termos textos curtos. Isso induzia em uma alta taxa de boletins similares apenas por terem palavras sobrepostas, por exemplo: "o relator acima qualificado informa que teve o seu celular roubado", e "O relator acima qualificado informa que teve sua bolsa roubada", embora os itens informados sejam completamente distintos, a representação gerada pelo TF-IDF apresenta alto grau de similaridade. Portanto, seria interessante explorar abordagens para lidar com esses casos e aprimorar a capacidade de identificar boletins semelhantes com base no conteúdo real dos boletins. Uma possibilidade seria realizar uma seleção de documentos baseada na quantidade de palavras contidas nos documentos, ignorando boletins com poucos detalhes.

Outra possibilidade é criar modelos que combinam um recurso léxico (ou um conjunto de recursos lexicais) com um recurso semântico (ou um conjunto de características semânticas) geralmente levam à geração de mais modelos precisos (Santos *et al.*, 2019). Isso ocorre porque essa abordagem permite que a similaridade seja medida simultaneamente em dois níveis diferentes (lexical + semântico). Outro ponto interessante seria a identificação do *modus operandi* baseado na sequência de ações descritas nos boletins de ocorrência, que poderia resultar em uma ferramenta poderosa para a polícia no auxílio a detecção e combate de crimes.

REFERÊNCIAS

AKBIK, A.; BERGMANN, T.; BLYTHE, D.; RASUL, K.; SCHWETER, S.; VOLLGRAF, R. Flair: An easy-to-use framework for state-of-the-art nlp. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**. [*S. l.*: *s. n.*], 2019. p. 54–59.

BOGGUST, A.; CARTER, B.; SATYANARAYAN, A. Embedding comparator: Visualizing differences in global structure and local neighborhoods via small multiples. In: **27th International Conference on Intelligent User Interfaces**. [*S. l.*: *s. n.*], 2022. p. 746–766.

BROMLEY ISABELLE GUYON, Y. L. E. S. J.; SHAH, R. Signature verification using a 'siamese' time delay neural network. In: **Advances in neural information processing systems.** [*S. l.*: *s. n.*], 1994. p. 737—-744.

CER, D.; YANG, Y.; KONG, S.-y.; HUA, N.; LIMTIACO, N.; JOHN, R. S.; CONSTANT, N.; GUAJARDO-CÉSPEDES, M.; YUAN, S.; TAR, C. *et al.* Universal sentence encoder. **arXiv preprint arXiv:1803.11175**, 2018.

CHOLLET, F. **Deep learning with Python**. [*S. l.*]: Simon and Schuster, 2021.

CHOPRA, S.; HADSELL, R.; LECUN, Y. Learning a similarity metric discriminatively, with application to face verification. In: **2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)**. [*S. l.*: *s. n.*], 2005. v. 1, p. 539–546 vol. 1.

CONNEAU, A.; KIELA, D.; SCHWENK, H.; BARRAULT, L.; BORDES, A. Supervised learning of universal sentence representations from natural language inference data. In: **Proceedings of the 2017 EMNLP**. [*S. l.*: *s. n.*], 2017. p. 670–680.

DO, P.; PHAM, P. W-kg2vec: a weighted text-enhanced meta-path-based knowledge graph embedding for similarity search. **Neural Computing and Applications**, Springer, v. 33, n. 23, p. 16533–16555, 2021.

FENG, F.; YANG, Y.; CER, D.; ARIVAZHAGAN, N.; WANG, W. Language-agnostic bert sentence embedding. **arXiv preprint arXiv:2007.01852**, 2020.

FINKELSTEIN, L.; GABRILOVICH, E.; MATIAS, Y.; RIVLIN, E.; SOLAN, Z.; WOLFMAN, G.; RUPPIN, E. Placing search in context: The concept revisited. In: **Proceedings of the 10th international conference on World Wide Web**. [*S. l.*: *s. n.*], 2001. p. 406–414.

FIRMIANO, A.; SILVA, T. L. C. D. Identifying duplicate police reports. In: **2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)**. [*S. l.*: *s. n.*], 2021. p. 244–247.

GHANNAY, S.; FAVRE, B.; ESTEVE, Y.; CAMELIN, N. Word embedding evaluation and combination. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. [*S. l.*: *s. n.*], 2016. p. 300–305.

HARANDI, M.; KUMAR, S. R.; NOCK, R. Siamese networks: A thing or two to know. **Data61, CSIRO**, 2017.

HARRIS, Z. S. Distributional structure. **Word**, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; SILVA, J.; ALUÍSIO, S. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: **Proceedings of the 11th STIL**. [*S. l.*: *s. n.*], 2017. p. 122–131.

HILL, F.; REICHART, R.; KORHONEN, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. **Computational Linguistics**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 41, n. 4, p. 665–695, 2015.

HJALTASON, G. R.; SAMET, H. Properties of embedding methods for similarity searching in metric spaces. **IEEE Transactions on Pattern Analysis and machine intelligence**, IEEE, v. 25, n. 5, p. 530–549, 2003.

IMTIAZ, Z.; UMER, M.; AHMAD, M.; ULLAH, S.; CHOI, G. S.; MEHMOOD, A. Duplicate questions pair detection using siamese malstm. **IEEE Access**, IEEE, v. 8, p. 21932–21942, 2020.

INOHARA, K.; UTSUMI, A. Jwsan: Japanese word similarity and association norm. **Language Resources and Evaluation**, Springer, v. 56, n. 1, p. 109–137, 2022.

IYYER, M.; MANJUNATHA, V.; BOYD-GRABER, J.; III, H. D. Deep unordered composition rivals syntactic methods for text classification. In: **Proceedings of the 53rd ACLWEB**. [*S. l.*: *s. n.*], 2015. p. 1681–1691.

KIROS, R.; ZHU, Y.; SALAKHUTDINOV, R. R.; ZEMEL, R.; URTASUN, R.; TORRALBA, A.; FIDLER, S. Skip-thought vectors. In: **Advances in neural information processing systems**. [*S. l.*: *s. n.*], 2015. p. 3294–3302.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: PMLR. **International conference on machine learning**. [*S. l.*], 2014. p. 1188–1196.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: **Advances in neural information processing systems**. [*S. l.*: *s. n.*], 2013. p. 3111–3119.

MILLER, G. A.; CHARLES, W. G. Contextual correlates of semantic similarity. **Language and cognitive processes**, Taylor & Francis, v. 6, n. 1, p. 1–28, 1991.

NAILI, M.; CHAIBI, A. H.; GHEZALA, H. H. B. Comparative study of word embedding methods in topic segmentation. **Procedia computer science**, Elsevier, v. 112, p. 340–349, 2017.

NAVIGLI, R.; MARTELLI, F. An overview of word and sense similarity. **Natural Language Engineering**, Cambridge University Press, v. 25, n. 6, p. 693–714, 2019.

NECULOIU, P.; VERSTEEGH, M.; ROTARU, M. Learning text similarity with siamese recurrent networks. In: **Proceedings of the 1st Workshop on Representation Learning for NLP**. [*S. l.*: *s. n.*], 2016. p. 148–157.

NEELAKANTAN TAO XU, R. P. A. R. J. M. H. J. T. Q. Y. N. T. J. W. K. C. H. J. H. P. S. B. P. T. E. N. G. S. G. K. D. S. F. P. S. K. H. M. T. T. K. T. S. J. J. P. W. L. W. A. **Text and Code Embeddings by Contrastive Pre-Training**. 2022.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 EMNLP**. [*S. l.*: *s. n.*], 2014. p. 1532–1543.

QIU, Y.; LI, H.; LI, S.; JIANG, Y.; HU, R.; YANG, L. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In: **Chinese computational linguistics and natural language processing based on naturally annotated big data**. [*S. l.*]: Springer, 2018. p. 209–221.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. **arXiv preprint arXiv:1908.10084**, 2019.

RODRIGUES, A. C.; MARCACINI, R. M. Sentence similarity recognition in portuguese from multiple embedding models. In: **2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)**. [*S. l.*: *s. n.*], 2022. p. 154–159.

RUBENSTEIN, H.; GOODENOUGH, J. B. Contextual correlates of synonymy. **Communications of the ACM**, ACM New York, NY, USA, v. 8, n. 10, p. 627–633, 1965.

SANTOS, J.; ALVES, A.; OLIVEIRA, H. G. Asapppy: a python framework for portuguese sts. In: **ASSIN@ STIL**. [*S. l.*: *s. n.*], 2019. p. 14–26.

SHAHMIRZADI, O.; LUGOWSKI, A.; YOUNGE, K. Text similarity in vector space models: a comparative study. In: IEEE. **2019 18th IEEE ICMLA**. [*S. l.*], 2019. p. 659–666.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **9th BRACIS**. [*S. l.*: *s. n.*], 2020.

TOSHEVSKA, M.; STOJANOVSKA, F.; KALAJDJIESKI, J. Comparative analysis of word embeddings for capturing word similarities. **arXiv preprint arXiv:2005.03812**, 2020.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. In: **Advances in neural information processing systems**. [*S. l.*: *s. n.*], 2017. p. 5998–6008.

VULIĆ, I.; BAKER, S.; PONTI, E. M.; PETTI, U.; LEVIANT, I.; WING, K.; MAJEWSKA, O.; BAR, E.; MALONE, M.; POIBEAU, T. *et al.* Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. **Computational Linguistics**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 46, n. 4, p. 847–897, 2020.

ZHAI, M.; TAN, J.; CHOI, J. Intrinsic and extrinsic evaluations of word embeddings. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [*S. l.*: *s. n.*], 2016. v. 30, n. 1.